SIAM J. MATH. ANAL.
Vol. 24, No. 1, pp. 1–22, January 1993

001

ENERGY ESTIMATES RELATING DIFFERENT LINEAR ELASTIC MODELS OF A THIN CYLINDRICAL SHELL I. THE MEMBRANE-DOMINATED CASE*

JYRKI PIILA† AND JUHANI PITKÄRANTA†

Abstract. Three different linear models describing the elastic deformation of a thin cylindrical shell are analyzed under a given smooth normal pressure distribution. A model problem with membrane-dominated state of deformation is considered. The models studied are (1) the standard three-dimensional model, (2) the classical shell model of Koiter–Sanders–Novozhilov, and (3) the asymptotic membrane theory of the shell. Estimates are derived relating the deformations fields according to different models in relative energy norm.

Key words. linear elasticity, energy estimates

AMS(MOS) subject classifications. 73C02, 73C20

1. Introduction. In this paper, which is the first part in a series of three papers, we compare three linear elastic models, describing the membrane-dominated deformation state of a finite, thin cylindrical shell. As a model problem we study a shell occupying, in Cartesian coordinates, the region

$$\Omega = \left\{ (x_1, x_2, x_3) \in R^3 \mid 0 < x_1 < H,\ R - \frac{t}{2} < \sqrt{x_2^2 + x_3^2} < R + \frac{t}{2} \right\}.$$

We will further choose $H = R = 1$ and assume that thickness t of the shell is small, i.e., $t << 1$. The shell is assumed to be loaded by a smoothly varying normal pressure distribution on the surface

$$\Gamma_+ = \left\{ (x_1, x_2, x_3) \in \overline{\Omega} \mid 0 < x_1 < 1,\ \sqrt{x_2^2 + x_3^2} = 1 + \frac{t}{2} \right\}.$$

The boundary at $x_1 = 0$ is assumed to be clamped, i.e., the displacements vanish at $x_1 = 0$, whereas the boundary at $x_1 = 1$ may be either clamped or free.

The three shell models to be considered are:

(1) The standard three-dimensional elastic model;

(2) The dimensionally reduced shell model of Koiter–Sanders–Novozhilov; and

(3) The asymptotic membrane theory of the shell.

We point out that shell problems fall, in general, into two categories depending on whether the membrane or bending deformations dominate. Here we concentrate on the membrane-dominated case. The bending-dominated case, closely related to the so-called inextensional shell theory, is the subject of Part II. In Part III we finally study another type of membrane-dominated situation referred to as the "soft" membrane case. In all three parts of the paper, the geometry of the shell is the same, and also the load is assumed to be of the same form as here, i.e., a smoothly varying normal pressure distribution. Thus only the boundary conditions and the shape of the load vary.

*Received by the editors November 19, 1990; accepted for publication (in revised form) February 24, 1992.

†Helsinki University of Technology, Institute of Mathematics, Otakaari 1, SF–02150 Espoo, Finland.

Here, in Part I, we take only the classical shell model of Koiter–Sanders–Novozhilov (as we have decided to call it; cf. [12], [20], [16]) as the representative of dimensionally reduced shell models. In Parts II and III we call this a model of Kirchhoff type and consider in addition what we call a model of Reissner–Mindlin type. The terminology here refers to the classical models of plate bending.

As is well known (cf. [14]), the displacement field $\underline{U}^{3D} = (U_1^{3D}, U_2^{3D}, U_3^{3D})$ minimizes the total energy

$$F^{3D}(\underline{U}) = \tfrac{1}{2}\mathcal{A}^{3D}(\underline{U},\underline{U}) - Q^{3D}(\underline{U}),$$

where the quadratic part $\mathcal{A}^{3D}(\underline{U},\underline{U})$ represents the deformation energy and the linear part $Q^{3D}(\underline{U})$ is the potential energy due to the external load. A dimensionally reduced shell model is obtained in this formulation by assuming that the displacements vary quadratically along directions normal to the midsurface of the shell. By dropping out certain "small" terms (see §3 below), the Koiter–Sanders–Novozhilov model is obtained. Finally, if the load is appropriately scaled in terms of t, we obtain a nontrivial and finite deformation state as $t \to 0$. In the present situation the limit state corresponds to the membrane theory of the shell, and it is obtained by letting the pressure distribution be proportional to t.

Let $\underline{U}^K$ stand for the displacement field corresponding to the shell model with positive thickness, and let $\underline{U}^0$ denote the membrane theory limit of $\underline{U}^K$. Our main results are estimates for $|||\underline{U}^{3D} - \underline{U}^K|||_{3D}$ and $|||\underline{U}^K - \underline{U}^0|||_{3D}$, where $|||\cdot|||_{3D}$ is the relative energy norm corresponding to the three-dimensional model. In particular, we show that in both the clamped-clamped and the clamped-free case

$$|||\underline{U}^{3D} - \underline{U}^K|||_{3D} = \mathcal{O}(t^{3/4} + \delta_\nu t^{1/2}), \tag{1.1a}$$

$$|||\underline{U}^K - \underline{U}^0|||_{3D} = \mathcal{O}(t^{1/4}), \tag{1.1b}$$

where ν, $0 \le \nu < \frac{1}{2}$, is the Poisson ratio of the material and

$$\delta_\nu = \begin{cases} 0 & \text{if } \nu = 0, \\ 1 & \text{otherwise.} \end{cases}$$

We also show that (1.1b) is the best possible estimate.

Estimates (1.1) are proved by splitting the displacement fields as $\underline{U}^{3D} = \underline{U}_A^{3D} + \underline{U}_B^{3D} + \underline{U}_C^{3D}$, and $\underline{U}^K = \underline{U}_A^K + \underline{U}_B^K$, where $\underline{U}_A^{3D}$ and $\underline{U}_A^K$ are both close to $\underline{U}^0$ and $\underline{U}_B^{3D}$, $\underline{U}_C^{3D}$ and $\underline{U}_B^K$ are residual fields arising because of the incompatibility of boundary conditions in different models. The residual fields satisfy homogeneous three-dimensional elastic/shell equations with certain inhomogeneous boundary conditions, see §4 below. In fact, we show that, with an appropriate splitting,

$$\begin{aligned} |||\underline{U}_A^{3D} - \underline{U}^0|||_{3D} &= \mathcal{O}(t), \\ |||\underline{U}_A^K - \underline{U}^0|||_{3D} &= \mathcal{O}(t), \\ |||\underline{U}_B^K|||_{3D} &= \mathcal{O}(t^{1/4}), \\ |||\underline{U}_B^{3D} - \underline{U}_B^K|||_{3D} &= \mathcal{O}(t^{3/4}), \\ |||\underline{U}_C^{3D}|||_{3D} &= \mathcal{O}(\delta_\nu t^{1/2}). \end{aligned}$$

The main difficulty in deriving energy estimates of the above type is to estimate the residual fields containing the (leading) boundary layer effects. So far we are only able to treat the simple geometry and boundary conditions considered. Yet, the results appear to be new even in this relatively simple case. For previous studies of the convergence of dimensionally reduced models for both plates and shells the reader is referred to [1]–[13], [15], [17]–[21]. Further references can be found, e.g., in [4], [6].

We point out that since $|||(\underline{U}_B^{3D}+\underline{U}_C^{3D})-\underline{U}_B^K|||_{3D}$ is asymptotically smaller than $|||\underline{U}_B^{3D}+\underline{U}_C^{3D}|||_{3D}$ in view of the estimates stated, this indicates that the (nonasymptotic) shell model does capture some of the leading boundary layers in the original three-dimensional formulation. This is in contrast to plate bending problems, where the Reissner–Mindlin model apparently does not exhibit such a superiority over the asymptotic Kirchhoff model [1], [2], [3].

Finally, let us mention that although the energy norm is certainly a "convenience norm" in our studies, the results are relevant from the numerical point of view since finite element approximations effectively generate best approximations in this norm; cf. [17], where the numerical aspects are discussed in the bending-dominated context.

The plan of the paper is as follows. Proceeding from the main notation in §2, we derive the dimensionally reduced shell models in §3 and also state some existence and convergence results of general nature. In §4 we prove convergence rate estimates (1.1) in the clamped-clamped case, and finally in §5 the clamped-free case is considered. In the Appendix we derive some regularity results for the displacement field $\underline{U}^K$ needed in estimating $|||\underline{U}^{3D}-\underline{U}^K|||_{3D}$. The need of such estimates is the main obstacle in extending the present results to more general situations.

Throughout the paper, the analysis is based on the energy principle or its variational form, i.e., we do not introduce the stresses at all. We depart here somewhat from the classical tradition where the stresses and the associated complementary energy principle usually play a vital role (cf. [13], [15], [21], [3], [11], [9]). In the subsequent Parts II and III we use the same analysis technique, except for one point in Part II where we still need the complementary energy principle.

2. Preliminaries. We summarize in this section the main notation to be used in the sequel. We work in a cylindrical coordinate system $(\alpha_1,\alpha_2,\alpha_3)$, where the shell occupies the region

$$\Omega=\left\{(\alpha_1,\alpha_2,\alpha_3)\in R^3 \mid (\alpha_1,\alpha_2)\in\omega,\ -\frac{t}{2}<\alpha_3<\frac{t}{2}\right\}.$$

Here (α_1,α_2) is a parametrization of the midsurface of the shell, with

$$\omega=\left\{(\alpha_1,\alpha_2)\in R^2 \mid 0<\alpha_1<1,\ -\pi<\alpha_2<\pi\right\}.$$

In the three-dimensional model of elasticity the strain tensor corresponding to a displacement field $\underline{V}$ is then defined as $\underline{\underline{e}}(\underline{V})=\{e_{ij}\}_{i,j=1}^3$, where (cf. [14])

$$\begin{aligned} e_{11}&=V_{1,1}, & e_{22}&=\chi(V_{2,2}+V_3),\\ e_{12}&=\tfrac12(\chi V_{1,2}+V_{2,1}), & e_{23}&=\tfrac12\big(V_{2,3}+\chi(V_{3,2}-V_2)\big),\\ e_{13}&=\tfrac12(V_{1,3}+V_{3,1}), & e_{33}&=V_{3,3}. \end{aligned}$$

Here the components V_i refer to the local orthonormal basis that corresponds to the curvilinear coordinate system; $V_{i,j}$ stands for $\partial V_i/\partial\alpha_j$ and $\chi=1/(1+\alpha_3)$. In the dimensionally reduced shell models the strain tensor is replaced by membrane strains

$\underline{\underline{\beta}} = \{\beta_{ij}\}_{i,j=1}^2$ and bending strains $\underline{\underline{\kappa}} = \{\kappa_{ij}\}_{i,j=1}^2$ depending only on (α_1, α_2). These have the expressions (cf. [16])

$$\begin{array}{lll} \beta_{11} = u_{,1}\,, & \beta_{12} = \tfrac{1}{2}(u_{,2} + v_{,1}\,), & \beta_{22} = v_{,2} + w, \\ \kappa_{11} = w_{,11}\,, & \kappa_{12} = w_{,12} - v_{,1}\,, & \kappa_{22} = w_{,22} - v_{,2}\,, \end{array}$$

where $\underline{u} = (u, v, w)$ is the displacement vector of the middle surface.

We introduce the bilinear forms

$$\begin{aligned} &\mathcal{A}^{3D}(\underline{U}, \underline{V}) \\ &\quad = D^{-1} \textstyle\int_\Omega \{\lambda\ tr\underline{\underline{e}}\ (\underline{U}) tr\underline{\underline{e}}(\underline{V}) + \mu \sum_{i,j=1}^3 e_{ij}(\underline{U}) e_{ij}(\underline{V})\} \chi^{-1}\, d\alpha_1\, d\alpha_2\, d\alpha_3, \\ &t^2 \mathcal{A}^K(\underline{u}, \underline{\tilde{v}}) + \mathcal{B}^K(\underline{u}, \underline{\tilde{v}}) \\ &\quad = t^2 \textstyle\int_\omega \{\nu\ tr\underline{\underline{\kappa}}\ tr\underline{\underline{\tilde{\kappa}}} + (1-\nu) \sum_{i,j=1}^2 \kappa_{ij}\ \tilde{\kappa}_{ij}\}\ d\alpha_1\ d\alpha_2 \\ &\qquad + 12 \textstyle\int_\omega \{\nu\ tr\underline{\underline{\beta}}\ tr\underline{\underline{\tilde{\beta}}} + (1-\nu) \sum_{i,j=1}^2 \beta_{ij}\ \tilde{\beta}_{ij}\}\ d\alpha_1\ d\alpha_2, \end{aligned}$$

corresponding to the three-dimensional and dimensionally reduced formulations, respectively. Here $tr\underline{\underline{e}} = e_{11} + e_{22} + e_{33}$, $tr\underline{\underline{\beta}} = \beta_{11} + \beta_{22}$, $tr\underline{\underline{\kappa}} = \kappa_{11} + \kappa_{22}$. Further, λ and μ are material parameters depending on the Young modulus $E > 0$ and the Poisson ratio ν, and D is a scaling factor. These are defined as

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}, \qquad \mu = \frac{E}{1+\nu}, \qquad D = \frac{Et}{12(1-\nu^2)}.$$

In the above notation, the scaled total energy to be minimized according to the three-dimensional model takes the form (cf. [14])

$$\begin{aligned} F^{3D}(\underline{U}) &= \tfrac{1}{2}\mathcal{A}^{3D}(\underline{U}, \underline{U}) - Q^{3D}(\underline{U}), \quad \text{where} \\ Q^{3D}(\underline{U}) &= \int_\omega f(\alpha_1, \alpha_2) \cdot U_3\left(\alpha_1, \alpha_2, \frac{t}{2}\right)\left(1 + \frac{t}{2}\right)\, d\alpha_1\, d\alpha_2. \end{aligned}$$

Here $f = D^{-1}F$ is the scaled load. We assume below that the actual load F is proportional to t, so f is independent of t. We see as well that this scaling assures (in the present membrane-dominated case) the existence of a nontrivial and finite limit deformation state as $t \to 0$.

Similarly, in the shell model of Koiter–Sanders–Novozhilov [16], the total energy is expressed as

$$\begin{aligned} F^K(\underline{u}) &= \tfrac{1}{2}\{t^2 \mathcal{A}^K(\underline{u}, \underline{u}) + \mathcal{B}^K(\underline{u}, \underline{u})\} - q(\underline{u}), \quad \text{where} \\ q(\underline{u}) &= \int_\omega f w\ d\alpha_1\ d\alpha_2, \end{aligned}$$

and f is as before. By taking the limit as $t \to 0$ we get formally the energy according to the membrane theory of the shell. This is denoted by F^0:

$$F^0(\underline{u}) = \tfrac{1}{2}\mathcal{B}^K(\underline{u}, \underline{u}) - q(\underline{u}). \tag{2.1}$$

The energy norms in the above three models are denoted by $|||\cdot|||_{3D}$, $|||\cdot|||_{K,t}$, and $|||\cdot|||_0$, respectively:

$$|||\underline{U}|||_{3D} = \sqrt{\mathcal{A}^{3D}(\underline{U},\underline{U})},$$
$$|||\underline{u}|||_{K,t} = \sqrt{t^2\mathcal{A}^K(\underline{u},\underline{u}) + \mathcal{B}^K(\underline{u},\underline{u})},$$
$$|||\underline{u}|||_0 = \sqrt{\mathcal{B}^K(\underline{u},\underline{u})}.$$

The corresponding energy spaces are denoted similarly by $\mathcal{U}^{3D}$, $\mathcal{U}^K$, and $\mathcal{U}^0$. Here

$$\mathcal{U}^{3D} = \{\underline{U} \in [H^1(\Omega)]^3 \mid \underline{U}(\cdot,-\pi,\cdot) = \underline{U}(\cdot,\pi,\cdot),\ \underline{U}(\alpha_1,\cdot,\cdot) = 0$$
$$\text{at } \alpha_1 \in \{0,\ 1\} \text{ in the clamped-clamped case or}$$
$$\text{at } \alpha_1 = 0 \text{ in the clamped-free case}\},$$

and

$$\mathcal{U}^K = \Bigg\{ (u,v,w) \in H^1(\omega) \times H^1(\omega) \times H^2(\omega) \mid$$
$$(u,v,w)(\cdot,-\pi) = (u,v,w)(\cdot,\pi),\ \frac{\partial w}{\partial \alpha_2}(\cdot,-\pi) = \frac{\partial w}{\partial \alpha_2}(\cdot,\pi),$$
$$(u,v,w)(\alpha_1,\cdot) = \frac{\partial w}{\partial \alpha_1}(\alpha_1,\cdot) = 0$$
$$\text{at } \alpha_1 \in \{0,\ 1\} \text{ in the clamped-clamped case or}$$
$$\text{at } \alpha_1 = 0 \text{ in the clamped-free case} \Bigg\},$$

where $H^p(\Omega)$ and $H^p(\omega)$ stand for the usual Sobolev spaces. To define $\mathcal{U}^0$, we note that if $\underline{u}$ is any smooth function vanishing at $\alpha_1 = 0$, then $\beta_{ij}(\underline{u}) = 0$, $i,j = 1,2$, implies $\underline{u} = \underline{0}$. Hence $|||\cdot|||_0$ defines a norm over such a set of functions and accordingly $\mathcal{U}^0$ is well defined as the closure of $\mathcal{U}^K$ with respect to $|||\cdot|||_0$. A partial characterization of $\mathcal{U}^0$ is given in Lemma 3.3 below.

In the following analysis we assume throughout that f is a restriction to ω of a smooth function $\tilde{f} = \tilde{f}(\alpha_1,\alpha_2)$ defined on R^2 and 2π-periodic in α_2. The set of such functions is denoted by $C^\infty_{\text{per}}(\overline{\omega})$. It would in fact suffice to assume a finite, sufficiently high degree of smoothness.

By C or c we denote various constants taking different values on different usage. The constants are independent of parameter t except when indicated explicitly. Further, we use the abbreviation $\mathcal{O}(\phi(t))$ for a quantity whose absolute value is bounded by $c(p)\phi(t)\|f\|_{p,\omega}$ for some finite p, where now and in the sequel $\|\cdot\|_{p,\omega}$ stands for the norm of the Sobolev space $H^p(\omega)$. Finally, the inner product of $L^2(\Omega)$ is denoted by $(\cdot,\cdot)$.

3. The shell models. The shell models to be considered can be derived from the basic assumption that the displacement field, expressed in terms of the curvilinear coordinates $(\alpha_1,\alpha_2,\alpha_3)$, is a quadratic function of α_3 for each (α_1,α_2). In particular we assume that

$$\begin{aligned} U_1(\alpha_1,\alpha_2,\alpha_3) &= u(\alpha_1,\alpha_2) - \alpha_3\theta_1(\alpha_1,\alpha_2), \\ U_2(\alpha_1,\alpha_2,\alpha_3) &= v(\alpha_1,\alpha_2) - \alpha_3\theta_2(\alpha_1,\alpha_2), \\ U_3(\alpha_1,\alpha_2,\alpha_3) &= w(\alpha_1,\alpha_2) + \alpha_3\psi_1(\alpha_1,\alpha_2) + \tfrac{1}{2}\alpha_3^2\psi_2(\alpha_1,\alpha_2). \end{aligned} \tag{3.1}$$

Displacement fields of this type were apparently first considered in [12]. Here $\underline{u} = (u, v, w)$ may be interpreted as the displacements of the midsurface of the shell along the coordinate axes; (θ_1, θ_2) are the so-called rotations, and (ψ_1, ψ_2) are auxiliary functions. The second-order terms relating U_1 and U_2 are skipped because they don't have any influence on the term $e_{33}(\underline{U}^{3D} - \underline{U})$, which alone causes the error of order $\mathcal{O}(\delta_\nu t^{1/2})$. (See Theorem 4.3.)

Upon making ansatz (3.1) in F^{3D}, expanding $1/(1+\alpha_3)$ into a Taylor series, and integrating with respect to α_3, we obtain

$$
\begin{aligned}
F^{3D}(\underline{U}) =& \frac{t^2}{2}\int_\omega \left\{ \nu (tr\underline{\underline{\kappa}})^2 + (1-\nu)\sum_{i,j=1}^{2} \kappa_{ij}^2 \right\} d\alpha_1\, d\alpha_2 \\
& + 6\int_\omega \left\{ \nu (tr\underline{\underline{\beta}})^2 + (1-\nu)\sum_{i,j=1}^{2} \beta_{ij}^2 \right\} d\alpha_1\, d\alpha_2 \\
& - q(\underline{u}) + R(\underline{U}),
\end{aligned}
\tag{3.2}
$$

where β_{ij} and q are as in §2,

$$
\kappa_{11} = \theta_{1,1}, \qquad \kappa_{12} = \tfrac{1}{2}(\theta_{1,2} + \theta_{2,1} - v_{,1}\,), \qquad \kappa_{22} = \theta_{2,2},
$$

and the remainder $R(\underline{U})$ is expanded as $R(\underline{U}) = \sum_{i=0}^{\infty} t^i R_i(\underline{U})$, where

$$
\begin{aligned}
R_0(\underline{U}) =& 3(1-\nu)\int_\omega \left\{ \rho_1^2 + \rho_2^2 + \frac{2(1-\nu)}{1-2\nu}\left(\psi_1 + \frac{\nu}{1-\nu} tr\underline{\underline{\beta}}\right)^2 \right\} d\alpha_1\, d\alpha_2, \\
R_1(\underline{U}) =& -\frac{1}{2}\int_\omega f\cdot(w + \psi_1)\, d\alpha_1\, d\alpha_2, \\
R_2(\underline{U}) =& \frac{1-\nu}{4}\int_\omega \left\{ \psi_{1,1}^2 + \psi_{1,2}^2 + \frac{2(1-\nu)}{1-2\nu}\left(\psi_2 - \frac{\nu}{1-\nu} tr\underline{\underline{\kappa}}\right)^2 - \frac{1}{1-\nu} f\left(\psi_1 + \frac{1}{2}\psi_2\right) \right. \\
& + \sum_{i,j=1}^{2} \beta_{ij} \left\{ \sum_{k,h=1}^{2} \left(c_{ijkh}\beta_{kh} + \tilde{c}_{ijkh}\kappa_{kh}\right) + c_{ij}\psi_1 + \tilde{c}_{ij}\psi_2 \right\} \\
& + \psi_1 \left\{ \sum_{i,j=1}^{2} \left(C_{ij}\beta_{ij} + \tilde{C}_{ij}\kappa_{ij}\right) + c_1\psi_1 + c_2\psi_2 \right\} \\
& \left. + c\rho_1^2 + C\rho_2^2 + \sum_{i=1}^{2} \rho_i(\hat{c}_{1i}\psi_{1,i} + \hat{c}_{2i}\psi_{2,i}) \right\} d\alpha_1\, d\alpha_2,
\end{aligned}
$$

etc., where $\rho_1 = -\theta_1 + w_{,1}$ and $\rho_2 = -\theta_2 + w_{,2} - v$.

The Koiter–Sanders–Novozhilov model is now obtained formally as the following.

Step 1. With $\underline{u} = \underline{u}^K$ fixed (to be defined in Step 3), minimize $R(\underline{U})$ approximately with respect to (θ_1, θ_2) by imposing the so-called Kirchhoff–Love constraints $\rho_1 = \rho_2 = 0$, i.e.,

$$
\theta_1^K = \frac{\partial w^K}{\partial \alpha_1}, \qquad \theta_2^K = \frac{\partial w^K}{\partial \alpha_2} - v^K.
\tag{3.3}
$$

Step 2. Minimize $R(\underline{U})$ approximately with respect to (ψ_1, ψ_2). Our choice is

$$\psi_1^K = -\frac{\nu}{1-\nu} tr\underline{\underline{\beta}}^K, \qquad \psi_2^K = \frac{\nu}{1-\nu} tr\underline{\underline{\kappa}}^K. \tag{3.4}$$

Step 3. Drop $R(\underline{U})$ and define $\underline{u}^K$ as the minimizer of the remaining part of the energy under constraints (3.3). So, $\underline{u}^K$ minimizes $F^K(\underline{u})$ as defined in §2.

In the above formalism, the asymptotic membrane theory is obtained by setting $t = 0$ in (3.2) and minimizing what is left. Thus, $\underline{u}^0$ minimizes $F^0(\underline{u})$ as defined by (2.1), $\theta_1^0 = w,_1^0$, $\theta_2^0 = w,_2^0 - v^0$, and $\psi_1^0 = -\nu tr\underline{\underline{\beta}}^0/(1-\nu)$. In this case we may set $\psi_2^0 = 0$.

Remark 3.1. Suppose that we do not drop the residual term $R_1(\underline{U})$ above, so that $\underline{\tilde{u}}^K$ is the minimizer of

$$t^2\mathcal{A}^K(\underline{u},\underline{u}) + \mathcal{B}^K(\underline{u},\underline{u}) - 2\tilde{q}(\underline{u}), \quad \text{where}$$

$$\tilde{q}(\underline{u}) = \int_\omega f\left(w\left(1+\frac{t}{2}\right) - \frac{\nu t}{2(1-\nu)} tr\underline{\underline{\beta}}\right) d\alpha_1\, d\alpha_2.$$

It is easy to verify that $|||\underline{\tilde{U}}^K - \underline{U}^K|||_{3D} = \mathcal{O}(t)$, so that this change has no effect on the main results (1.1a) and (1.1b).

Regarding the solvability of the above dimensionally reduced problems, we note first that, due to the well-known Korn inequality,

$$t^2\mathcal{A}^K(\underline{u},\underline{u}) + \mathcal{B}^K(\underline{u},\underline{u}) \geq ct^2\left(||u||_{1,\omega}^2 + ||v||_{1,\omega}^2 + ||w||_{2,\omega}^2\right), \qquad \underline{u} \in \mathcal{U}^K,$$

the existence and uniqueness of $\underline{u}^K$ is guaranteed by the Riesz representation theorem. Also $\underline{u}^0$ exists and is unique by the same argument, for we have the following.

LEMMA 3.1. *For any $f \in C^\infty_{\text{per}}(\overline{\omega})$, q is a bounded linear functional on $\mathcal{U}^0$.*

Proof. Let $\underline{u} \in \mathcal{U}^K$. Integrating by parts, we get

$$|q(\underline{u})| = \left|\int_\omega \left(f\beta_{22} - 2I_1[f,_2]\beta_{12} + I_2[f,_{22}]\beta_{11}\right) d\alpha_1\, d\alpha_2\right| \leq C||f||_{2,\omega}|||\underline{u}|||_0,$$

where $I_n[g](\alpha_1,\alpha_2)$ is defined as

$$I_n[g](\alpha_1,\alpha_2) = \int_1^{\alpha_1}\int_1^{x_1}\cdots\int_1^{x_{n-1}} g(x_n,\alpha_2)\, dx_n\, dx_{n-1}\cdots dx_1.$$

Since $\mathcal{U}^K$ is dense in $\mathcal{U}^0$, the assertion follows. □

We can characterize $\underline{u}^0$ as follows.

LEMMA 3.2. *For any $f \in C^\infty_{\text{per}}(\overline{\omega})$, $\underline{u}^0 = (u^0, v^0, w^0) \in [C^\infty_{\text{per}}(\overline{\omega})]^3$ and $u^0 = v^0 = 0$ at the clamped ends of the cylinder.*

Note that since $(u,v,w) \to \left\{||u||_{1,\omega}^2 + ||v||_{1,\omega}^2 + ||w||_{L_2(\omega)}^2\right\}^{1/2}$ is a stronger norm than $|||\cdot|||_0$, we cannot impose any boundary conditions on w^0 at the clamped ends. In the proof of Lemma 3.2 we need the following characterization of the energy space $\mathcal{U}^0$.

LEMMA 3.3.

$$\mathcal{U}^0 \cap [C^\infty_{\text{per}}\ (\overline{\omega})]^3 = \{\underline{u} \in [C^\infty_{\text{per}}\ (\overline{\omega})]^3 \mid u = v = 0 \textit{ at the clamped ends of the cylinder}\}.$$

Proof. Let $\{\underline{u}_n\}_{n=1}^{\infty} = \{(u_n, v_n, w_n)\}_{n=1}^{\infty}$ be a sequence in $\mathcal{U}^0 \cap [C^{\infty}_{\text{per}}(\bar{\omega})]^3$ such that $\underline{u}_n \to \underline{u} \in \mathcal{U}^0 \cap [C^{\infty}_{\text{per}}(\bar{\omega})]^3$ in the norm $|||\cdot|||_0$ and such that $u_n = v_n = 0$ for all n at the clamped ends. Then, assuming a Fourier expansion of $\underline{u}_n$ with respect to α_2 as in the Appendix, it follows that each Fourier component of u_n and v_n converges to the corresponding component of u, respectively, v in $H^1(0,1)$. Accordingly, each Fourier component of u and v vanishes at the clamped ends and hence so do u and v. On the other hand, if $\underline{u} \in [C^{\infty}_{\text{per}}(\bar{\omega})]^3$ is given such that $u = v = 0$ at the clamped ends, then we can find a sequence $\{w_n\}_{n=1}^{\infty}$ of smooth functions such that $w_n = \partial w_n/\partial \alpha_1 = 0$ at the clamped ends and such that $w_n \to w$ in $L_2(\omega)$. Then $(u, v, w_n) \to (u, v, w)$ in the norm of $\mathcal{U}^0$, and accordingly, $\underline{u} \in \mathcal{U}^0$. □

Proof of Lemma 3.2. We simply solve $\underline{u}^0$ explicitly. First note that $\underline{u}^0$ must satisfy the Euler equations

$$\begin{aligned} \beta_{11,1} + \nu\beta_{22,1} + (1-\nu)\beta_{12,2} &= 0, \\ \nu\beta_{11,2} + \beta_{22,2} + (1-\nu)\beta_{12,1} &= 0, \\ 12\nu\beta_{11} + 12\beta_{22} &= f, \end{aligned}$$

and if the end at $\alpha_1 = 1$ is free, also the natural boundary conditions

$$\beta_{11} + \nu\beta_{22} = \beta_{12} = 0 \quad \text{at} \quad \alpha_1 = 1.$$

If $\underline{u}^0$ is smooth, we can impose the boundary condition $u^0(0,\cdot) = v^0(0,\cdot) = 0$ by Lemma 3.3. Then $\underline{u}^0$ must be of the form

$$\begin{aligned} u^0 &= \tilde{c}\{I_3[f_{,22}] - \nu I_1[f]\} - \alpha_1^2\xi' + \alpha_1\eta, \\ v^0 &= -\tilde{c}\{I_4[f_{,222}] + (2+\nu)I_2[f_{,2}]\} + \frac{\alpha_1^3}{3}\xi'' + 4(1+\nu)\alpha_1\xi - \frac{\alpha_1^2}{2}\eta', \\ w^0 &= -\nu u_{,1} - v_{,2} + \frac{f}{12}, \end{aligned}$$

where now $\tilde{c} = 1/\big(12(1-\nu^2)\big)$,

$$I_n[g](\alpha_1, \alpha_2) = \int_0^{\alpha_1}\int_0^{x_1}\cdots\int_0^{x_{n-1}} g(x_n, \alpha_2)\, dx_n\, dx_{n-1}\cdots dx_1,$$

and ξ and η are functions of α_2 only. In the clamped-clamped case ξ and η are determined by solving the system

$$\begin{aligned} &\xi' - \eta = g_1, \quad \tfrac{1}{3}\xi'' + 4(1+\nu)\xi - \tfrac{1}{2}\eta' = g_2, \quad \text{where} \\ &g_1 = \tilde{c}\{I_3[f_{,22}](1,\cdot) - \nu I_1[f](1,\cdot)\}, \\ &g_2 = \tilde{c}\{I_4[f_{,222}](1,\cdot) + (2+\nu)I_2[f_{,2}](1,\cdot)\}, \end{aligned}$$

under the boundary conditions $\xi(-\pi) = \xi(\pi)$, $\xi'(-\pi) = \xi'(\pi)$, $\eta(-\pi) = \eta(\pi)$. In the clamped-free case ξ and η are defined so as to satisfy the mentioned natural boundary conditions at $\alpha_1 = 1$. Obviously, ξ and η are in both cases restrictions to $[-\pi, \pi]$ of smooth 2π-periodic functions defined for all $\alpha_2 \in R$. We have thus found a solution of the desired type to the limit variational problem that is in $\mathcal{U}^0$ by Lemma 3.3. It follows from Lemma 3.1 that this solution is unique in the sense of the norm of $\mathcal{U}^0$. □

So far we have defined $\underline{u}^0$ only as a formal limit of $\underline{u}^K$ as $t \to 0$. We show next that $\underline{u}^K$ actually converges to $\underline{u}^0$ in the norm $|||\cdot|||_{K,t}$. The proof is a restatement of the classical regularization argument of Tihonov [22].

THEOREM 3.1. $|||\underline{u}^K - \underline{u}^0|||_{K,t} \to 0$ *as* $t \to 0$.

Proof. We note first that for all $\underline{u} \in \boldsymbol{\mathcal{U}}^K$,

$$F^K(\underline{u}) = F^0(\underline{u}) + \frac{t^2}{2}\mathcal{A}^K(\underline{u},\underline{u}) = \frac{1}{2}|||\underline{u} - \underline{u}^0|||_0^2 - \frac{1}{2}|||\underline{u}^0|||_0^2 + \frac{t^2}{2}\mathcal{A}^K(\underline{u},\underline{u}).$$

Let $\{\underline{u}_n\}_{n=1}^{\infty} \subset \boldsymbol{\mathcal{U}}^K$ be a sequence such that $|||\underline{u}_n - \underline{u}^0|||_0 \to 0$ as $n \to \infty$. Then since $F^K(\underline{u}^K) \leq F^K(\underline{u}_n)$ for all n, it follows that

$$|||\underline{u}^K - \underline{u}^0|||_0^2 + t^2\mathcal{A}^K(\underline{u}^K,\underline{u}^K) \leq |||\underline{u}_n - \underline{u}^0|||_0^2 + t^2\mathcal{A}^K(\underline{u}_n,\underline{u}_n).$$

We can now fix first n and then t so that the right-hand side is arbitrarily small and hence the assertion follows easily. □

4. Convergence rate estimates, the clamped-clamped case. We start by splitting the displacement fields $\underline{U}^K$ and $\underline{U}^{3D}$. First $\underline{U}^K$ is rewritten as $\underline{U}^K = \underline{U}_A^K + \underline{U}_B^K$, where $\underline{U}_A^K$ is defined in terms of $\underline{u}_A^K = (u_A^K,\ v_A^K,\ w_A^K)$ according to (3.1), (3.3), and (3.4), and further $\underline{u}_A^K$ minimizes F^K under periodic boundary condition at $\alpha_2 \in \{-\pi,\ \pi\}$ and under the inhomogeneous boundary conditions

$$\underline{u}(\alpha_1,\cdot) = \underline{u}^0(\alpha_1,\cdot), \qquad \frac{\partial w}{\partial \alpha_1}(\alpha_1,\cdot) = \frac{\partial w^0}{\partial \alpha_1}(\alpha_1,\cdot) \quad \text{at } \alpha_1 \in \{0,\ 1\}.$$

The remaining part $\underline{U}_B^K$ is then likewise defined in terms of $\underline{u}_B^K$ according to (3.1), (3.3), and (3.4), where $\underline{u}_B^K$ minimizes the homogeneous functional $\underline{u} \to \frac{1}{2}\{t^2\mathcal{A}^K(\underline{u},\underline{u}) + \mathcal{B}^K(\underline{u},\underline{u})\}$ under the periodic boundary condition at $\alpha_2 \in \{-\pi,\ \pi\}$ and under the boundary conditions

$$\underline{u}(\alpha_1,\cdot) = -\underline{u}^0(\alpha_1,\cdot), \qquad \frac{\partial w}{\partial \alpha_1}(\alpha_1,\cdot) = -\frac{\partial w^0}{\partial \alpha_1}(\alpha_1,\cdot) \quad \text{at } \alpha_1 \in \{0,\ 1\}.$$

Next, let $\gamma = 1 + \frac{t}{2}$. (This factor is needed essentially to compensate for the dissymmetry of the loading with respect to the shell midsurface; see the proof of Theorem 4.1 below.) We split the three-dimensional displacement field $\underline{U}^{3D}$ corresponding to a given load f as $\underline{U}^{3D} = \underline{U}_A^{3D} + \underline{U}_B^{3D} + \underline{U}_C^{3D}$, where $\underline{U}_A^{3D}$ minimizes F^{3D} under the periodic boundary condition at $\alpha_2 \in \{-\pi,\ \pi\}$ and under the inhomogeneous boundary conditions

$$\underline{U}(\alpha_1,\cdot,\cdot) = \gamma\underline{U}_A^K(\alpha_1,\cdot,\cdot), \qquad \alpha_1 \in \{0,\ 1\}.$$

The remaining parts $\underline{U}_B^{3D}$ and $\underline{U}_C^{3D}$ minimize the homogeneous functional $\underline{U} \to \frac{1}{2}\mathcal{A}^{3D}(\underline{U},\underline{U})$ under the boundary conditions

$$\underline{U}(\alpha_1,\cdot,\cdot) = \gamma\underline{U}_B^K(\alpha_1,\cdot,\cdot) \quad \text{and}$$

$$\underline{U}(\alpha_1,\cdot,\cdot) = \left(0,\ 0,\ \gamma\frac{\nu}{1-\nu}\left(\alpha_3 tr\underline{\underline{\beta}}(\underline{u}^K) - \frac{1}{2}\alpha_3^2 tr\underline{\underline{\kappa}}(\underline{u}^K)\right)\right)(\alpha_1,\cdot,\cdot),$$

respectively, at $\alpha_1 \in \{0,\ 1\}$.

Our aim is to prove that

$$|||\underline{U}_A^K - \underline{U}^0|||_{3D} = \mathcal{O}(t), \qquad |||\underline{u}_A^K - \underline{u}^0|||_{K,t} = \mathcal{O}(t^2), \tag{4.1}$$

$$|||\underline{U}_B^K|||_{3D} = \mathcal{O}(t^{1/4}), \qquad |||\underline{u}_B^K|||_{K,t} = \mathcal{O}(t^{1/4}), \tag{4.2}$$

$$|||\underline{U}_A^{3D} - \underline{U}_A^K|||_{3D} = \mathcal{O}(t), \tag{4.3}$$

$$|||\underline{U}_B^{3D} - \underline{U}_B^K|||_{3D} = \mathcal{O}(t^{3/4}), \tag{4.4}$$

$$|||\underline{U}_C^{3D}|||_{3D} = \mathcal{O}(\delta_\nu t^{1/2}). \tag{4.5}$$

The main results (1.1) are obviously consequences of these estimates. We will also prove that (4.2) and thus (1.1b) is the best possible estimate. Note that in view of (4.1)–(4.5), the leading irregularities of $\underline{U}^{3D}$ and $\underline{U}^K$ due to the clamped boundaries are contained in $\underline{U}_B^{3D} + \underline{U}_C^{3D}$ and $\underline{U}_B^K$, respectively. Moreover, we conclude that

$$\frac{|||(\underline{U}_B^{3D} + \underline{U}_C^{3D}) - \underline{U}_B^K|||_{3D}}{|||\underline{U}_B^{3D} + \underline{U}_C^{3D}|||_{3D}} = \mathcal{O}(t^{1/2} + \delta_\nu t^{1/4}),$$

so $\underline{U}_B^K$ is a "good" approximation of $\underline{U}_B^{3D} + \underline{U}_C^{3D}$ in the sense of relative energy norm.

Estimates (4.1) and (4.2) follow immediately from Lemmas A.1 and A.2 of the Appendix. Note that the second parts of (4.1) and (4.2) sharpen the result of Theorem 3.1. It thus suffices to prove (4.3) through (4.5). We start by proving (4.4). By Lemma A.2, $\underline{u}_B^K \in [C^\infty_{\text{per}}(\overline{\omega})]^3$, and for any multi-index $\tau = (\tau_1, \tau_2)$,

$$\begin{aligned} &\|D^\tau u_B^K\| = t^{-\sigma}\mathcal{O}(t^{1/4} + t^{(3-2\tau_1)/4}),\\ (4.6)\qquad &\|D^\tau v_B^K\| = t^{-\sigma}\mathcal{O}(t^{1/4} + t^{(5-2\tau_1)/4}),\\ &\|D^\tau w_B^K\| = t^{-\sigma}\mathcal{O}(t^{(1-2\tau_1)/4}), \end{aligned}$$

where $\sigma = 0$ if $\|\cdot\| = \|\cdot\|_{L_2(\omega)}$ and $\sigma = \frac{1}{4}$ if $\|\cdot\| = \|\cdot\|_{L_\infty(\omega)}$. Applying these regularity estimates we can prove the following.

THEOREM 4.1. $|||\underline{U}_B^{3D} - \underline{U}_B^K||| = \mathcal{O}(t^{3/4})$.

Proof. Let $\tilde{\underline{U}}_B^{3D}$ be the solution of the variational problem $\mathcal{A}^{3D}(\tilde{\underline{U}}_B^{3D}, \underline{V}) = 0$ for all $\underline{V} \in \boldsymbol{\mathcal{U}}^{3D}$ with boundary conditions $\tilde{\underline{U}}_B^{3D}(\alpha_1, \cdot, \cdot) = \underline{U}_B^K(\alpha_1, \cdot, \cdot)$ at $\alpha_1 \in \{0,\ 1\}$. Then by linearity $\underline{U}_B^{3D} = \gamma \tilde{\underline{U}}_B^{3D}$, and it follows from the triangle inequality that

$$|||\underline{U}_B^{3D} - \underline{U}_B^K|||_{3D} \le \gamma |||\tilde{\underline{U}}_B^{3D} - \underline{U}_B^K|||_{3D} + \frac{t}{2}|||\underline{U}_B^K|||_{3D}.$$

Since $|||\underline{U}_B^K|||_{3D} = \mathcal{O}(1)$ at most, it obviously suffices to prove the assertion with $\underline{U}_B^{3D}$ replaced by $\tilde{\underline{U}}_B^{3D}$. By a simple calculation,

$$\begin{aligned} e_{11}(\underline{U}_B^K) &= \beta_{11} - \alpha_3\kappa_{11}, & e_{12}(\underline{U}_B^K) &= \chi\beta_{12} - \chi^2\alpha_3\kappa_{12} + R_{12},\\ e_{13}(\underline{U}_B^K) &= R_{13}, & e_{22}(\underline{U}_B^K) &= \beta_{22} - \alpha_3\kappa_{22} + R_{22},\\ e_{23}(\underline{U}_B^K) &= R_{23}, & e_{33}(\underline{U}_B^K) &= -\frac{\nu}{1-\nu}\,(tr\underline{\underline{\beta}} - \alpha_3 tr\underline{\underline{\kappa}}), \end{aligned}$$

where throughout this proof $\beta_{ij} = \beta_{ij}(\underline{u}_B^K)$, $\kappa_{ij} = \kappa_{ij}(\underline{u}_B^K)$. Applying (4.6), the remainder terms $R_{ij} = R_{ij}(\underline{U}_B^K)$ can be easily estimated as

$$(4.7)\qquad \|R_{ij}\|_{L_2(\Omega)} = \mathcal{O}(t^{5/4}).$$

We note that since $\underline{u}_B^K$ minimizes $t^2\mathcal{A}^K(\underline{u}, \underline{u}) + \mathcal{B}^K(\underline{u}, \underline{u})$, we have the Euler equations

$$(4.8a)\qquad \beta_{11,1} + \nu\beta_{22,1} + (1-\nu)\beta_{12,2} = 0,$$

$$(4.8b)\qquad \nu\beta_{11,2} + \beta_{22,2} + (1-\nu)\beta_{12,1} - \frac{t^2}{12}\cdot K_1 = 0,$$

$$(4.8c)\qquad 12\nu\beta_{11} + 12\beta_{22} + t^2\cdot K_2 = 0,$$

where

$$K_1 = \nu\kappa_{11,2} + 2(1-\nu)\kappa_{12,1} + \kappa_{22,2},$$
$$K_2 = \kappa_{11,11} + \nu\kappa_{22,11} + 2(1-\nu)\kappa_{12,12} + \nu\kappa_{11,22} + \kappa_{22,22}.$$

Let $\underline{V} \in \mathcal{U}^{3D}$. Integrating by parts and applying (4.8a) and (4.8b), we have

(4.9)
$$\begin{aligned}\mathcal{A}^{3D}(\underline{U}_B^K, \underline{V}) = \frac{12}{t}\Big\{&(\nu\beta_{11} + \beta_{22}, V_3) - \big(\kappa_{11} + \nu\kappa_{22}, V_{1,1}(\alpha_3 + \alpha_3^2)\big) \\ &- \left((1-\nu)\kappa_{12}, \frac{\alpha_3 V_{1,2}}{(1+\alpha_3)^2} + \frac{\alpha_3 V_{2,1}}{1+\alpha_3}\right) \\ &- \big(\nu\kappa_{11} + \kappa_{22}, \alpha_3(V_{2,2} + V_3)\big)\Big\} \\ &- t(K_1, V_2) + R_1(\underline{V}), \qquad |R_1(\underline{V})| \le C(f)\cdot t^{3/4}\cdot |||\underline{V}|||_{3D},\end{aligned}$$

where the estimate follows from (4.6) and (4.7). Next, again integrating by parts and recalling (4.6), we find that

$$\begin{aligned}&\frac{12}{t}\int_\Omega (\nu\beta_{11} + \beta_{22}) V_3 \, d\alpha_1\, d\alpha_2 \,\partial_3\left(\alpha_3 + \frac{t}{2}\right) \\ &= \int_\omega (12\nu\beta_{11} + 12\beta_{22}) V_3\left(\cdot, \frac{t}{2}\right) d\alpha_1\, d\alpha_2 \;+ R_2(\underline{V}), \\ &|R_2(\underline{V})| \le C(f)\cdot t^{5/4}\cdot |||\underline{V}|||_{3D}.\end{aligned} \tag{4.10}$$

The remaining terms on the right side of (4.9) can be handled in a similar manner. As the details get here more complicated, we collect the required results first in the following.

LEMMA 4.1. *Let $\underline{V} \in \mathcal{U}^{3D}$. Then*

$$\begin{aligned}&-\frac{12}{t}\big(\kappa_{11} + \nu\kappa_{22}, V_{1,1}(\alpha_3 + \alpha_3^2)\big) \\ &= t^2\int_\omega (\kappa_{11,11} + \nu\kappa_{22,11}) V_3\left(\cdot, \frac{t}{2}\right) d\alpha_1\, d\alpha_2 + R_3(\underline{V}),\end{aligned} \tag{4.11}$$

$$\begin{aligned}&-\frac{12}{t}\left((1-\nu)\kappa_{12}, \frac{\alpha_3 V_{1,2}}{(1+\alpha_3)^2} + \frac{\alpha_3 V_{2,1}}{1+\alpha_3}\right) \\ &= t^2\int_\omega 2(1-\nu)\kappa_{12,12} V_3\left(\cdot, \frac{t}{2}\right) d\alpha_1\, d\alpha_2 + R_4(\underline{V}),\end{aligned} \tag{4.12}$$

$$\begin{aligned}&-\frac{12}{t}\big(\nu\kappa_{11} + \kappa_{22}, \alpha_3(V_{2,2} + V_3)\big) - t(K_1, V_2) \\ &= t^2\int_\omega (\nu\kappa_{11,22} + \kappa_{22,22}) V_3\left(\cdot, \frac{t}{2}\right) d\alpha_1\, d\alpha_2 + R_5(\underline{V}),\end{aligned} \tag{4.13}$$

where $|R_j(\underline{V})| \le C(f)\cdot t^{3/4}\cdot |||\underline{V}|||_{3D}, j = 3, 4, 5.$

End of the proof of Theorem 4.1. It follows from (4.8c) and (4.9)–(4.13) that for all $\underline{V} \in \mathcal{U}^{3D}$, $\mathcal{A}^{3D}(\underline{U}_B^K - \underline{\tilde{U}}_B^{3D}, \underline{V}) = \sum_{j=1}^5 R_j(\underline{V})$. By the above estimates, the assertion of Theorem 4.1 follows. □

Thus it suffices to prove Lemma 4.1.
Proof of (4.11).

$$\begin{aligned}
&-\frac{12}{t}(\kappa_{11}+\nu\kappa_{22}, V_{1,1}(\alpha_3+\alpha_3^2))\\
&\quad=\frac{12}{t}\int_\Omega(\kappa_{11,1}+\nu\kappa_{22,1})V_1\,d\alpha_1\,d\alpha_2\,\partial_3\left(\frac{\alpha_3^2}{2}-\frac{t^2}{8}\right)+r_1\\
&\quad=-\frac{12}{t}\left(\kappa_{11,1}+\nu\kappa_{22,1},(2e_{13}(\underline{V})-V_{3,1})\left(\frac{\alpha_3^2}{2}-\frac{t^2}{8}\right)\right)+r_1\\
&\quad=-\frac{12}{t}\left(\kappa_{11,11}+\nu\kappa_{22,11},V_3\left(\frac{\alpha_3^2}{2}-\frac{t^2}{8}\right)\right)+r_1+r_2\\
&\quad=-\frac{12}{t}\int_\Omega(\kappa_{11,11}+\nu\kappa_{22,11})V_3\,d\alpha_1\,d\alpha_2\,\partial_3\left(\frac{\alpha_3^3}{6}-\frac{\alpha_3t^2}{8}-\frac{t^3}{24}\right)+r_1+r_2\\
&\quad=t^2\int_\omega(\kappa_{11,11}+\nu\kappa_{22,11})V_3\left(\cdot,\frac{t}{2}\right)\,d\alpha_1\,d\alpha_2+r_1+r_2+r_3.
\end{aligned}$$

By (4.6), $|R_3(\underline{V})|=\left|\sum_{i=1}^3 r_i(\underline{V})\right|\le C(f)\cdot t^{3/4}\cdot|||\underline{V}|||$ and the assertion follows.
Proof of (4.12).

$$\begin{aligned}
&-\frac{12(1-\nu)}{t}\left(\kappa_{12},\frac{\alpha_3V_{1,2}}{(1+\alpha_3)^2}+\frac{\alpha_3V_{2,1}}{1+\alpha_3}\right)\\
&=-\frac{12(1-\nu)}{t}\int_\Omega\kappa_{12}\left(\frac{V_{1,2}}{(1+\alpha_3)^2}+\frac{V_{2,1}}{1+\alpha_3}\right)\,d\alpha_1\,d\alpha_2\,\partial_3\left((1+\alpha_3)^2\left(\frac{\alpha_3^2}{2}-\frac{t^2}{8}\right)\right)+r_1\\
&=-\frac{12(1-\nu)}{t}\left(\kappa_{12,2}\left(V_{1,3}-\frac{2V_1}{1+\alpha_3}\right)+\kappa_{12,1}\left(V_{2,3}(1+\alpha_3)-V_2\right),\left(\frac{\alpha_3^2}{2}-\frac{t^2}{8}\right)\right)+r_1\\
&=I+II+r_1.
\end{aligned}$$

Taking into account that $V_{1,3}=2e_{13}-V_{3,1}$ and also that

$$||V_1||_{L_2(\Omega)}\le||e_{11}(\underline{V})||_{L_2(\Omega)},$$

we have integrating by parts;

$$\begin{aligned}
I&=-\frac{12(1-\nu)}{t}\left(\kappa_{12,12},V_3\left(\frac{\alpha_3^2}{2}-\frac{t^2}{8}\right)\right)+r_2\\
&=t^2\int_\omega(1-\nu)\kappa_{12,12}V_3\left(\cdot,\frac{t}{2}\right)\,d\alpha_1\,d\alpha_2+r_2+r_3,
\end{aligned}$$

where the last equality follows in the same way as in the previous proof. Similarly,

$$\begin{aligned}
II=&-\frac{12}{t}(1-\nu)\left(\kappa_{12,1},(2(1+\alpha_3)e_{23}(\underline{V})-V_{3,2})\left(\frac{\alpha_3^2}{2}-\frac{t^2}{8}\right)\right)\\
=&t^2\int_\omega(1-\nu)\kappa_{12,12}V_3\left(\cdot,\frac{t}{2}\right)\,d\alpha_1\,d\alpha_2+r_4.
\end{aligned}$$

By (4.6), $|R_4(\underline{V})|=\left|\sum_{i=1}^4 r_i(\underline{V})\right|\le C(f)\cdot t^{5/4}\cdot|||\underline{V}|||$, and also this part of the proof is complete.

Proof of (4.13).

$$
\begin{aligned}
&-\frac{12}{t}\left(\nu\kappa_{11}+\kappa_{22},\alpha_3(V_{2,2}+V_3)\right)\\
&\quad=-\frac{12}{t}\left(\nu\kappa_{11}+\kappa_{22},(V_{2,2}+V_3)\left(\alpha_3+2\alpha_3^2-\frac{t^2}{4}\right)\right)+r_1\\
&\quad=I+II+r_1,
\end{aligned}
$$

where

$$
\begin{aligned}
I=&-\frac{12}{t}\left(\nu\kappa_{11}+\kappa_{22},V_{2,2}\left(\alpha_3+2\alpha_3^2-\frac{t^2}{4}\right)\right)\\
=&\frac{12}{t}\int_\Omega(\nu\kappa_{11,2}+\kappa_{22,2})\frac{V_2}{1+\alpha_3}\;d\alpha_1\;d\alpha_2\;\partial_3\left((1+\alpha_3)^2\left(\frac{\alpha_3^2}{2}-\frac{t^2}{8}\right)\right)\\
=&t^2\int_\omega(\nu\kappa_{11,22}+\kappa_{22,22})V_3\left(\cdot,\frac{t}{2}\right)d\alpha_1\;d\alpha_2+r_2.
\end{aligned}
$$

Rewriting $-t(K_1,V_2)=-t(\nu\kappa_{11}+\kappa_{22},V_3)+r_3$, we have

$$
\begin{aligned}
II-t(K_1,V_2)=&-\frac{12}{t}\int_\Omega(\nu\kappa_{11}+\kappa_{22})V_3\;d\alpha_1\;d\alpha_2\;\partial_3\left(\frac{\alpha_3^2}{2}-\frac{t^2}{8}+\frac{2\alpha_3^3}{3}-\frac{t^2\alpha_3}{4}\right)\\
&-t\int_\Omega(\nu\kappa_{11}+\kappa_{22})V_3\;d\alpha_1\;d\alpha_2\;\partial_3(\alpha_3)+r_3=r_3+r_4.
\end{aligned}
$$

Defining $R_5(\underline{V})=\sum_{i=1}^4 r_i$, we again conclude from (4.6) that

$$
|R_5(\underline{V})|=\left|\sum_{i=1}^4 r_i(\underline{V})\right|\le C(f)\cdot t^{5/4}\cdot|||\underline{V}|||_{3D},
$$

and so, combining the above identities, the last part of Lemma 4.1 is proved. □

To prove (4.3) we cite first from Lemma A.1 that $\underline{u}_A^K\in[C^\infty_{\text{per}}(\overline{\omega})]^3$ and that

$$
||u_A^K||_{5,\omega}+||v_A^K||_{6,\omega}+||w_A^K||_{4,\omega}=\mathcal{O}(1). \tag{4.14}
$$

THEOREM 4.2. $|||\underline{U}_A^{3D}-\underline{U}_A^K|||_{3D}=\mathcal{O}(t)$.

Proof. First of all, we can replace $\underline{U}_A^{3D}$ by $\underline{\tilde{U}}_A^{3D}=\frac{1}{\gamma}\underline{U}_A^{3D}$ so that $\underline{\tilde{U}}_A^{3D}-\underline{U}_A^K$ is kinematically admissible. Proceeding then exactly in the same way as in the previous proof, we find that $\mathcal{A}^{3D}(\underline{U}_A^K-\underline{\tilde{U}}_A^{3D},\underline{V})=\delta(\underline{V})$ for all $\underline{V}\in\mathcal{U}^{3D}$. Using estimates (4.14) we can easily verify that in this case $|\delta(\underline{V})|\le C(f)\cdot t\cdot|||\underline{V}|||_{3D}$. Hence the assertion follows. □

THEOREM 4.3. $|||\underline{U}_C^{3D}|||_{3D}=\mathcal{O}(\delta_\nu t^{1/2})$.

Proof. Let $\underline{U}=(U_1,U_2,U_3)$, where $U_1=U_2=0$ and

$$
U_3=\gamma\frac{\nu}{1-\nu}\left(\alpha_3 tr\underline{\underline{\beta}}(\underline{u}^K)-\frac{1}{2}\alpha_3^2 tr\underline{\underline{\kappa}}(\underline{u}^K)\right)e^{-\alpha_1(1-\alpha_1)/t}.
$$

Then by the definition of $\underline{U}_C^{3D}$, and by (4.6) and (4.14),

$$
|||\underline{U}_C^{3D}|||_{3D}\le|||\underline{U}|||_{3D}=\mathcal{O}(\delta_\nu t^{1/2}).
$$

□

Remark 4.1. Let us show that (4.2), and thus (1.1b), is the best possible estimate. We assume that $f \equiv 1$. In that case the solution of (4.8) is of the form $\big(u_B^K(\alpha_1), 0, w_B^K(\alpha_1)\big)$. After a simple computation, we get

$$\begin{aligned} w_B^K = & C_0 + C_1 e^{-A\alpha_1}\cos A\alpha_1 + C_2 e^{-A\alpha_1}\sin A\alpha_1 \\ & + C_3 e^{A\alpha_1}\cos A\alpha_1 + C_4 e^{A\alpha_1}\sin A\alpha_1, \end{aligned} \tag{4.15}$$

where $A = \big(3(1-\nu)\big)^{1/4} t^{-1/2}$. Taking into account the desired boundary conditions and also the constraint

$$\int_0^1 \big(12 w_B^K + (w_B^K)^{(4)}\big)\, d\alpha_1 = 0,$$

which follows from (4.8c), we get five linearly independent equations for the coefficients C_i in (4.15). Using symbolic calculus (MATHEMATICA) we obtain the exact estimates

$$C_0 = \mathcal{O}(\delta_\nu t^{1/2}),\quad C_1 = \mathcal{O}(1),\quad C_2 = \mathcal{O}(1),\quad C_3 = \mathcal{O}(e^{-A}),\quad C_4 = \mathcal{O}(e^{-A}). \tag{4.16}$$

Further, u_B^K can be solved from (4.8a):

$$u_B^K = -\nu \int_0^{\alpha_1} w_B^K(\tau)\, d\tau + \nu\alpha_1 \int_0^1 w_B^K(\tau)\, d\tau. \tag{4.17}$$

By (4.15), (4.16), and (4.17),

$$\lim_{t\to 0} t^{-3/4} \|e_{22}(\underline{U}_B^K)\|_{L_2(\Omega)} \sim \lim_{t\to 0} t^{-1/4} \|\beta_{22}\|_{L_2(\omega)} \sim \mathcal{O}(1),$$

and hence (4.2) cannot be improved.

Next, let us define the interior domain of the cylinder,

$$\Omega_\delta := \big\{(\alpha_1, \alpha_2, \alpha_3) \in \Omega \mid \delta < \alpha_1 < 1 - \delta\,\big\},\qquad 0 < \delta < \tfrac{1}{2},$$

and also the corresponding energy norm

$$|||\underline{U}|||_\delta^2 = D^{-1} \int_{\Omega_\delta} \left\{ \lambda\big(tr\ \underline{\underline{e}}(\underline{U})\big)^2 + \mu \sum_{i,j} e_{ij}(\underline{U})^2 \right\} (1+\alpha_3)\, d\alpha_1\, d\alpha_2\, d\alpha_3.$$

It follows from (4.15), (4.16), and (4.17) that

$$|||\underline{U}_B^K|||_\delta = \mathcal{O}(t^{1/4} e^{-\delta t^{-1/2}} + \delta_\nu t^{1/2}). \tag{4.18}$$

This shows that $\underline{U}_B^K$ is a true bondary layer only if $\nu = 0$. Anyhow, it follows from (4.1)–(4.5) and (4.18) that

$$|||\underline{U}^K - \underline{U}^L|||_\delta \le C(f)(t + \delta_\nu t^{1/2}),\qquad \delta \ge t^{1/2}\ln(t^{-3/4}), \tag{4.19}$$

so the interior convergence rate is faster than the global one.

5. The clamped-free case. In this last section we consider the case where the boundary at $\alpha_1 = 1$ is free. The splitting of $\underline{U}^K$ and $\underline{U}^{3D}$ is otherwise similar to the previous chapter, but now the boundary values are only fixed at the clamped end.

We are going to prove that (4.1)–(4.5) still hold with one exception, namely, that the second part of (4.1) must be rewritten as

$$|||\underline{u}_A^K - \underline{u}^0|||_{K,t} = \mathcal{O}(t^{5/4}). \tag{4.1$'$}$$

Again (4.1′) and (4.2) are consequences of Lemmas A.1 and A.2, respectively. To prove the rest, we proceed exactly in the same way as in §4. Using the natural boundary conditions

$$\begin{aligned} &\beta_{11}^K + \nu\beta_{22}^K = 0, \quad && 6\beta_{12}^K - t^2\kappa_{12}^K = 0, \\ &\kappa_{11}^K + \nu\kappa_{22}^K = 0, \quad && \kappa_{11,1}^K + \nu\kappa_{22,1}^K + 2(1-\nu)\kappa_{12,2}^K = 0 \end{aligned} \tag{5.1}$$

at the free end, we find that

$$\mathcal{A}^{3D}(\underline{U}_X^K - \underline{\tilde{U}}_X^{3D}, \underline{V}) = \delta_X(\underline{V}) - \int_\Gamma \underline{g}^X \cdot \underline{V}(1,\cdot,\cdot)\, d\alpha_2\, d\alpha_3$$

for all $\underline{V} \in \boldsymbol{\mathcal{U}}^{3D}$, where X stands for either A or B, $\Gamma = \{(1, \alpha_2, \alpha_3) \in \overline{\Omega}\}$,

$$\underline{g}^X = \frac{12}{t}(1-\nu)\left\{0,\ \kappa_{12}^X(1,\alpha_2)\left(\alpha_3 + 2\alpha_3^2 - \frac{t^2}{4}\right),\ \kappa_{12,2}^X(1,\alpha_2)\left(\frac{\alpha_3^2}{2} - \frac{t^2}{8}\right)\right\},$$

and where $\delta_X(\underline{V}) = \sum_{j=1}^5 R_j(\underline{V})$ consists of the same residual terms as in the proofs of Theorems 4.1 and 4.2. It follows from Lemma A.1 that $\underline{u}_A^K \in [C_{\text{per}}^\infty(\overline{\omega})]^3$ and that

$$\begin{aligned} \|\, D^\tau u_A^K\|_{L_2(\omega)} &= \mathcal{O}(1 + t^{(7-2\tau_1)/4}), \\ \|\, D^\tau v_A^K\|_{L_2(\omega)} &= \mathcal{O}(1 + t^{(9-2\tau_1)/4}), \\ \|D^\tau w_A^K\|_{L_2(\omega)} &= \mathcal{O}(1 + t^{(5-2\tau_1)/4}) \end{aligned} \tag{5.2}$$

for every multi-index $\tau = (\tau_1, \tau_2)$. Using (5.2) and (4.6), which is valid also in the clamped-free case, we get, as in the previous section,

$$|\delta_A(\underline{V})| \le C(f)\cdot t \cdot |||\underline{V}|||_{3D}, \qquad |\delta_B(\underline{V})| \le C(f)\cdot t^{3/4}\cdot |||\underline{V}|||_{3D}. \tag{5.3}$$

We need finally an estimate for the displacement field $\underline{U}^X \in U^{3D}$ defined so that

$$\mathcal{A}^{3D}(\underline{U}^X, \underline{V}) = \int_\Gamma \underline{g}^X \cdot \underline{V}(1,\cdot,\cdot)\, d\alpha_2\, d\alpha_3, \quad \underline{V} \in \boldsymbol{\mathcal{U}}^{3D}. \tag{5.4}$$

In the Appendix (see Lemma A.4) we prove that

$$|||\underline{U}^X|||_{3D} = \mathcal{O}(t). \tag{5.5}$$

After these preliminaries, the main result of this chapter can be stated.

THEOREM 5.1. *Theorems* 4.1, 4.2, *and* 4.3 *hold also in the clamped-free case.*

Proof. Observing that

$$\mathcal{A}^{3D}(\underline{U}_X^K + \underline{U}^X - \underline{\tilde{U}}_X^{3D}, \underline{U}_X^K + \underline{U}^X - \underline{\tilde{U}}_X^{3D}) = \delta_X(\underline{U}_X^K + \underline{U}^X - \underline{\tilde{U}}_X^{3D}),$$

the assertions of Theorems 4.1 and 4.2 follow from (5.3), (5.5), and the triangle inequality. Theorem 4.3 can be proved in the same way as before. □

Remark 5.1. Let us assume that $f \equiv 1$ as we did in Remark 4.1. Solving (4.8) under the relevant natural boundary conditions at $\alpha_1 = 1$, we find that

$$u_B^K(\alpha_1) = -\nu \int_0^{\alpha_1} w_B^K(\tau)\, d\tau,$$

and w_B^K is of the form (4.15), where this time

$$C_0 = 0, \quad C_1 = \mathcal{O}(1), \quad C_2 = \mathcal{O}(1), \quad C_3 = \mathcal{O}(e^{-2A}), \quad C_4 = \mathcal{O}(e^{-2A}).$$

It turns out easily that (4.2) cannot be improved, but now $\underline{U}_B^K$ is a pure boundary layer. Estimate (4.19) can this time be rewritten as

$$|||\underline{U}^K - \underline{U}^0|||_\delta = \mathcal{O}(t), \qquad \delta \geq t^{1/2}\ln(t^{-3/4}),$$

where now $\Omega_\delta = \{\underline{\alpha} \in \Omega |\ \alpha_1 > \delta\}$ and $|||\cdot|||_\delta$ is the corresponding norm.

Appendix. Here we (1) study the regularity properties of the displacement field $\underline{u}^K$ as defined according to the Koiter–Sanders–Novozhilov shell model, and (2) prove estimate (5.5) related to the three-dimensional model. The analysis is based on Fourier expansions.

First consider the dimensionally reduced model. Here we expand the load and the displacement field as

$$f = \sum_{k=0}^{\infty} \big(f^{ck}(\alpha_1)\cos k\alpha_2 + f^{sk}(\alpha_1)\sin k\alpha_2\big),$$

$$(u,v,w) = \sum_{k=0}^{\infty} \big(u^{ck}(\alpha_1)\cos k\alpha_2,\ v^{ck}(\alpha_1)\sin k\alpha_2,\ w^{ck}(\alpha_1)\cos k\alpha_2\big)$$
$$+ \sum_{k=1}^{\infty} \big(u^{sk}(\alpha_1)\sin k\alpha_2, -v^{sk}(\alpha_1)\cos k\alpha_2,\ w^{sk}(\alpha_1)\sin k\alpha_2\big).$$

It is easy to check that the energy is then split orthogonally as

$$F^K(\underline{u}) = \sum_{k=0}^{\infty} c_k F_{ck}^K(u^{ck}, v^{ck}, w^{ck}) + \sum_{k=1}^{\infty} c_k F_{sk}^K(u^{sk}, v^{sk}, w^{sk}),$$

where $c_0 = 2\pi$, $c_k = \pi$ for $k \geq 1$, and

$$F_{ck}^K(\underline{u}) = \tfrac{1}{2}\{t^2\mathcal{A}_k^F(\underline{u},\underline{u}) + \mathcal{B}_k^F(\underline{u},\underline{u})\} - q_{ck}^F(\underline{u}),$$
$$F_{sk}^K(\underline{u}) = \tfrac{1}{2}\{t^2\mathcal{A}_k^F(\underline{u},\underline{u}) + \mathcal{B}_k^F(\underline{u},\underline{u})\} - q_{sk}^F(\underline{u}),$$

where further

$$q_{ck}^F(\underline{u}) = \int_0^1 f^{ck} w\, d\alpha_1, \qquad q_{sk}^F(\underline{u}) = \int_0^1 f^{sk} w\, d\alpha_1,$$

$$\mathcal{A}_k^F(\underline{u},\underline{u}) = \int_0^1 \left\{ \nu(tr\underline{\underline{\kappa}}^k)^2 + (1-\nu)\sum_{i,j=1}^{2}(\kappa_{ij}^k)^2 \right\} d\alpha_1,$$

$$\mathcal{B}_k^F(\underline{u},\underline{u}) = 12\int_0^1 \left\{ \nu(tr\underline{\underline{\beta}}^k)^2 + (1-\nu)\sum_{i,j=1}^{2}(\beta_{ij}^k)^2 \right\} d\alpha_1 \quad \text{with}$$

$$\beta_{11}^k = u', \quad \beta_{12}^k = \tfrac{1}{2}(-ku + v'), \quad \beta_{22}^k = kv + w,$$
$$\kappa_{11}^k = w'', \quad \kappa_{12}^k = -kw' - v', \quad \kappa_{22}^k = -k^2 w - kv.$$

Due to the orthogonality of the above, splitting it suffices to consider one Fourier component of $\underline{u}^K$ at a time in proving regularity estimates. Thus assume $f = g(\alpha_1)\cos k\alpha_2$ or $f = g(\alpha_1)\sin k\alpha_2$ define the energy space

$$\boldsymbol{\mathcal{U}} = \{\underline{u} \in [H^1(I)]^2 \times H^2(I) \mid \underline{u} = w' = 0 \text{ at the clamped endpoints}\},$$

where $\underline{u} = (u, v, w)$ and $I = (0,\ 1)$, and consider the variational problem.

Find $\underline{u}^k \in \boldsymbol{\mathcal{U}}$ such that

$$\text{(A.1)} \qquad t^2 \mathcal{A}_k^F(\underline{u}^k, \underline{\tilde{v}}) + \mathcal{B}_k^F(\underline{u}^k, \underline{\tilde{v}}) = \int_I g\tilde{w}\, d\alpha_1 \qquad \text{for all } \underline{\tilde{v}} \in \boldsymbol{\mathcal{U}}.$$

Below we denote by $|||\cdot|||_k$ the corresponding energy norm and by $||\cdot||_{s,I}$, $s \geq 0$, the norm of the Sobolev space $H^s(I)$.

Next, we denote by $\underline{u}_0^k$ the limit of $\underline{u}^k$ as $t \to 0$ and split then $\underline{u}^k$ as $\underline{u}^k = \underline{u}_A^k + \underline{u}_B^k$, where $\underline{u}_A^k$ satisfies (A.1) together with the inhomogeneous boundary conditions

$$u_A^k = u_0^k, \quad v_A^k = v_0^k, \quad w_A^k = w_0^k, \quad (w_A^k)' = (w_0^k)'$$

at the the clamped ends. Then $\underline{u}_B^k$ satisfies (A.1) with $g = 0$, together with the boundary conditions

$$u_B^k = -u_0^k, \quad v_B^k = -v_0^k, \quad w_B^k = -w_0^k, \quad (w_B^k)' = -(w_0^k)'$$

at the clamped ends.

In the four lemmas below we use the abbreviation $\mathcal{O}(\phi(t))$ for a quantity bounded by $c(p,k)\phi(t)||g||_{p,I}$ for some finite p, where the dependence on parameter k is algebraic, i.e., there exists $m = m(p) \in N$ such that, $c(p,k) \leq c(p)k^m$.

LEMMA A.1. *For any* $s \geq 0$,

$$\begin{aligned} ||u_A^k - u_0^k||_{s,I} &= t^{-\sigma}\mathcal{O}(t^2 + t^{(5-s)/2}), \\ ||v_A^k - v_0^k||_{s,I} &= t^{-\sigma}\mathcal{O}(t^2 + t^{(6-s)/2}), \\ ||w_A^k - w_0^k||_{s,I} &= t^{-\sigma}\mathcal{O}(t^{(4-s)/2}), \end{aligned}$$

where $\sigma = 0$ *in the clamped-clamped case and* $\sigma = \frac{3}{4}$ *in the clamped-free case.*

LEMMA A.2. *For any* $s \geq 0$,

$$\begin{aligned} ||u_B^k\ || &= t^{-\sigma}\mathcal{O}(t^{1/4} + t^{(3-2s)/4}), \\ ||v_B^k\ || &= t^{-\sigma}\mathcal{O}(t^{1/4} + t^{(5-2s)/4}), \\ ||w_B^k|| &= t^{-\sigma}\mathcal{O}(t^{(1-2s)/4}), \end{aligned}$$

where $\sigma = 0$ *if* $||\cdot|| = ||\cdot||_{s,I}$ *and* $\sigma = \frac{1}{4}$ *if* $||u|| = \sum_{m \leq s} ||D^m u||_{L_\infty(I)}$.

LEMMA A.3. *Assume the clamped-free case. Then for* $0 \leq m \leq 3$,

$$(u_B^k)^{(m)}(1) = \mathcal{O}(1), \quad (v_B^k)^{(m)}(1) = \mathcal{O}(1), \quad (w_B^k)^{(m)}(1) = \mathcal{O}(1).$$

Proof of Lemma A.1. Setting $\underline{u} = \underline{u}_A^k - \underline{u}_0^k \in \boldsymbol{\mathcal{U}}$, we have integrating by parts that

(A.2)

$$t^2\mathcal{A}_k^F(\underline{u},\underline{\tilde{v}}) + \mathcal{B}_k^F(\underline{u},\underline{\tilde{v}}) = t^2\left\{\int_I (h_1\tilde{v} + h_2\tilde{w})\,d\alpha_1 + \delta\big(A\tilde{w}'(1) + B\tilde{w}(1) + C\tilde{v}(1)\big)\right\}$$

for all $\underline{\tilde{v}} = (\tilde{u},\tilde{v},\tilde{w}) \in \boldsymbol{\mathcal{U}}$, where h_1 and h_2 are t–independent smooth functions, A, B, and C t-independent scalars and either $\delta = 0$ (clamped-clamped case) or $\delta = 1$ (clamped-free case). Noting that

$$\|\tilde{w}\|_{L_2(I)} \le ck^2|||\underline{\tilde{v}}|||_k, \tag{A.3a}$$

$$|\tilde{v}(1)| \le \|\tilde{v}\|_{1,I} \le ck|||\underline{\tilde{v}}|||_k \tag{A.3b}$$

for all $\underline{\tilde{v}} \in \boldsymbol{\mathcal{U}}$, we get the energy estimate

$$|||\underline{u}|||_k = \mathcal{O}(t^2) \quad \text{in the clamped-clamped case.} \tag{A.4}$$

Further,

$$\|\tilde{w}\|_{2,I} \le ct^{-1}|||\underline{\tilde{v}}|||_k, \tag{A.5a}$$

and thus by interpolation,

$$\|\tilde{w}\|_{1,I} \le ckt^{-1/2}|||\underline{\tilde{v}}|||_k. \tag{A.5b}$$

Next, writing

$$\tilde{w}^{(j)}(1) = \int_{1-\epsilon}^{1} \partial\left(\frac{\tau-1+\epsilon}{\epsilon}\tilde{w}^{(j)}(\tau)\right) = \int_{1-\epsilon}^{1}\left\{\frac{1}{\epsilon}\tilde{w}^{(j)}(\tau) + \frac{\tau-1+\epsilon}{\epsilon}\tilde{w}^{(j+1)}(\tau)\right\}\,d\tau$$

and using the Cauchy–Schwarz inequality, we get

$$|\tilde{w}^{(j)}(1)|^2 \le \frac{2}{\epsilon}\int_I [\tilde{w}^{(j)}(\tau)]^2\,d\tau + \frac{2\epsilon}{3}\int_I [\tilde{w}^{(j+1)}(\tau)]^2\,d\tau. \tag{A.6}$$

Substituting $\epsilon = t^{1/2}$ and recalling (A.3a), (A.5a), and (A.5b), we have

$$|\tilde{w}(1)| \le ck^2t^{-1/4}|||\underline{\tilde{v}}|||_k, \tag{A.7a}$$

$$|\tilde{w}'(1)| \le ckt^{-3/4}|||\underline{\tilde{v}}|||_k. \tag{A.7b}$$

By (A.2), (A.3a), (A.3b), (A.7a), and (A.7b) we get as a counterpart to (A.4),

$$|||\underline{u}|||_k = \mathcal{O}(t^{5/4}) \quad \text{in the clamped-free case.} \tag{A.8}$$

By (A.3), (A.4), (A.5), and (A.8) we further get the a priori estimates

$$\|u\|_{1,I} = \mathcal{O}(t^{2-\sigma}), \quad \|v\|_{1,I} = \mathcal{O}(t^{2-\sigma}), \quad \|w\|_{s,I} = \mathcal{O}(t^{(4-s-2\sigma)/2}), \quad 0 \le s \le 2,$$

with σ as in Lemma A.1. Applying these results in the Euler equations

$$u'' = \tfrac{1}{2}k^2(1-\nu)u - \tfrac{1}{2}k(1+\nu)v' - \nu w', \tag{A.9a}$$

$$v'' = c\left\{t^2h_1 - 6k\big((1+\nu)u' + 2(kv+w)\big) + kt^2\big((2-\nu)w'' - k^2w - kv\big)\right\}, \tag{A.9b}$$

$$w^{(4)} = h_2 + 2k^2w'' - k^4w + k(2-\nu)v'' - k^3v - 12t^{-2}(\nu u' + kv + w), \tag{A.9c}$$

where $c = -1/\big(2(1-\nu)(3+t^2)\big)$, we get the a posteriori estimates for $||u||_{2,I}$, $||v||_{2,I}$ and $||w||_{4,I}$. Finally, differentiating in (A.9), the proof is completed by induction and interpolation. □

Proof of Lemma A.2. Let $\underline{u} \in \boldsymbol{\mathcal{U}}$ be such that $u = v = 0$ and $w = -w_0^k(\alpha_1)\cdot\rho(\alpha_1)$, where in the clamped-clamped case

$$\rho(\alpha_1) = e^{-t^{-1/2}\alpha_1(1-\alpha_1)}\big(\cos t^{-1/2}\alpha_1(1-\alpha_1) + \sin t^{-1/2}\alpha_1(1-\alpha_1)\big),$$

and in the clamped-free case

$$\rho(\alpha_1) = e^{-t^{-1/2}\alpha_1}(\cos t^{-1/2}\alpha_1 + \sin t^{-1/2}\alpha_1).$$

Then $t^2\mathcal{A}_k^F(\underline{u},\underline{u}) + \mathcal{B}_k^F(\underline{u},\underline{u}) = \mathcal{O}(t^{1/2})$ and since $\underline{u}_B^k$ minimizes in $\boldsymbol{\mathcal{U}}$ this functional, we get the a priori estimates

$$||u_B^k||_{1,I} = \mathcal{O}(t^{1/4}), \quad ||v_B^k||_{1,I} = \mathcal{O}(t^{1/4}), \quad ||w_B^k||_{s,I} = \mathcal{O}(t^{(1-2s)/4}), \quad 0 \le s \le 2.$$

Applying these results in Euler equations (A.9), where now and in the sequel $h_1 = h_2 = 0$, the assertion with $||\cdot|| = ||\cdot||_{s,I}$ follows. Noting that (A.6) holds with 1 and $\tilde{w}$ replaced by any $\alpha_1 \in I$ and smooth function g, the second part of the assertion follows by recalling the Sobolev-norm estimates already obtained. □

Proof of Lemma A.3. In this proof we use the notation $(u,v,w) \leftarrow (u_B^k, v_B^k, w_B^k)$. Furthermore, we make change of variable $\alpha_1 \leftarrow 1-\alpha_1$ so as to switch the clamped and free ends.

By (A.6) and Lemma A.2, $w(0) = \mathcal{O}(1)$. By the same argument $u^{(j)}(0) = \mathcal{O}(1)$ for $j = 0,1$ and $v^{(j)}(0) = \mathcal{O}(1)$ for $j = 0,1,2$. Taking into account equations (A.9), the natural boundary conditions

$$\text{(A.10)} \qquad w''(0) = \nu\big(k^2w(0) + kv(0)\big), \qquad w^{(3)}(0) = (2-\nu)\big(k^2w'(0) + kv'(0)\big),$$

and the equation

$$u' = -\nu w - k(1-\nu)\int_0^{\alpha_1} \beta_{12}(\tau)\,d\tau - k\nu v$$

obtained from (A.9a) using the boundary condition $\beta_{11}(0) + \nu\beta_{22}(0) = 0$, it is enough to prove that $w'(0) = \mathcal{O}(1)$. Substituting the last equation into (A.9c), we get the equation of the form

$$\text{(A.11)} \qquad w^{(4)} + 4A^4w = g, \quad \text{where } A = \big(3(1-\nu^2)t^{-2} + \tfrac{1}{4}k^4\big)^{1/4} = \mathcal{O}(t^{-1/2})$$

and where by Lemma A.2, g satisfies the estimate;

$$\max_{\alpha_1\in[0,1]}\big\{|g(\alpha_1)| + |g'(\alpha_1)|\big\} = \mathcal{O}(t^{-2}).$$

By (A.11) and (A.10), we have

$$\begin{aligned}
\text{(A.12)} \qquad w(\alpha_1) &= a^{(3)}(\alpha_1)\cdot w(0) + a''(\alpha_1)\cdot w'(0) + b(\alpha_1), \quad \text{where}\\
a(\alpha_1) &= \frac{1}{4A^3}\big(\sin A\alpha_1\cdot\cosh A\alpha_1 - \cos A\alpha_1\cdot\sinh A\alpha_1\big),\\
b(\alpha_1) &= \int_0^{\alpha_1} g(\alpha_1-\tau)a(\tau)\,d\tau + a'(\alpha_1)\cdot w''(0) + a(\alpha_1)\cdot w'''(0).
\end{aligned}$$

By (A.10), $w''(0) = \mathcal{O}(1)$ and, also applying (A.6), $w'''(0) = \mathcal{O}(t^{-1/2})$ at most. Taking into account that $w(1) = -w_0^k(1)$ and $w'(1) = -(w_0^k)'(1)$, it follows from (A.12) that

$$\begin{bmatrix} w(0) \\ w'(0) \end{bmatrix} = -2\cosh^{-2} A \begin{bmatrix} a^{(3)}(1) & -a''(1) \\ -a^{(4)}(1) & a^{(3)}(1) \end{bmatrix} \cdot \begin{bmatrix} b(1) + w_0^k(1) \\ b'(1) + (w_0^k)'(1) \end{bmatrix}. \tag{A.13}$$

Noting that $a^{(4)}(\alpha_1) = -4A^4 a(\alpha_1)$ and also that $a^{(3)}(0) = 1$, we have, integrating by parts,

$$b(1) = -\frac{g(0)a^{(3)}(1)}{4A^4} + \mathcal{O}(t^{1/2}e^A).$$

On the other hand, $b'(1) = g(0)a(1) + \mathcal{O}(e^A)$, and thus

$$-a^{(4)}(1)b(1) + a^{(3)}(1)b'(1) = 4A^4 a(1)b(1) + a^{(3)}(1)b'(1) = \mathcal{O}(e^{2A}).$$

The assertion follows now from (A.13). □

We prove finally the energy estimate (5.5) for the solution of (5.4). In the same way as above, the solution of (5.4) can be expanded into Fourier series

$$\begin{aligned} \underline{U}^X = & \sum_{k=0}^{\infty} \big(U_1^{ck}(\alpha_1,\alpha_3)\cos k\alpha_2,\ U_2^{ck}(\alpha_1,\alpha_3)\sin k\alpha_2,\ U_3^{ck}(\alpha_1,\alpha_3)\cos k\alpha_2\big) \\ & + \sum_{k=1}^{\infty} \big(U_1^{sk}(\alpha_1,\alpha_3)\sin k\alpha_2,\ -U_2^{sk}(\alpha_1,\alpha_3)\cos k\alpha_2,\ U_3^{sk}(\alpha_1,\alpha_3)\sin k\alpha_2\big). \end{aligned}$$

Each component $(U_1^{\cdot k}, U_2^{\cdot k}, U_3^{\cdot k}) \in \mathcal{U}^{2D}$ satisfies the variational equation

$$\begin{aligned} & \mathcal{A}^{2D}(\underline{U}^X, \underline{V}) \\ & = D^{-1} \int_{\tilde\Omega} \left\{ \lambda\, tr\underline{\underline{e}}(\underline{U}^X) tr\underline{\underline{e}}(\underline{V}) + \mu \sum_{i,j=1}^{3} e_{ij}(\underline{U}^X) e_{ij}(\underline{V}) \right\} (1+\alpha_3)\, d\alpha_1\, d\alpha_3 \\ & = \int_{\tilde\Gamma} \underline{g}^X(\alpha_3) \cdot \underline{V}(1,\alpha_3)\, d\alpha_3, \qquad \underline{V} \in \mathcal{U}^{2D}, \end{aligned}$$

where $\tilde\Gamma = (-t/2, t/2)$, $\tilde\Omega = I \times \tilde\Gamma$, $\underline{\underline{e}}$ is the strain tensor corresponding to $\underline{V}$,

$$\begin{aligned} e_{11} &= V_{1,1}, & e_{22} &= \chi(kV_2 + V_3), \\ e_{12} &= \tfrac{1}{2}\,(-\chi k V_1 + V_{2,1}), & e_{23} &= \tfrac{1}{2}\,\big(V_{2,3} - \chi(kV_3 + V_2)\big), \\ e_{13} &= \tfrac{1}{2}\,(V_{1,3} + V_{3,1}), & e_{33} &= V_{3,3}, \end{aligned}$$

and finally

$$\underline{g}^X(\alpha_3) = \frac{12(1-\nu)\kappa_{12}^X(1)}{t} \left(0,\ \alpha_3 + 2\alpha_3^2 - \frac{t^2}{4},\ k\left(\frac{\alpha_3^2}{2} - \frac{t^2}{8}\right)\right).$$

Above, κ_{12}^X is the Fourier component of κ_{12}^X corresponding to $f^{\cdot k}$. Due to the orthogonality of the basis functions, (5.5) holds if we prove the following.

LEMMA A.4. $|||\underline{U}^X|||_{2D} = \mathcal{O}(t)$.

Proof. It follows from Lemmas A.1 (in the spirit of (A.6)) and A.3 that independently of X,

$$(\text{A.14}) \qquad \|\hat{g}_2^X\|_{L_2(\tilde{\Gamma})} = \mathcal{O}(t^{1/2}), \qquad \|g_3^X\|_{L_2(\tilde{\Gamma})} = \mathcal{O}(t^{3/2}).$$

Further, we obtain easily the estimates

$$(\text{A.15a}) \qquad \|V_2\|_{L_2(\tilde{\Omega})} \le \|V_{2,1}\|_{L_2(\tilde{\Omega})} \le ckt^{1/2}|||\underline{V}|||_{2D},$$

$$(\text{A.15b}) \qquad \|V_3\|_{L_2(\tilde{\Omega})} \le ck^2t^{1/2}|||\underline{V}|||_{2D}$$

for all $\underline{V} \in \mathcal{U}^{2D}$. Then, if $\underline{V} \in \mathcal{U}^{2D}$, let

$$\hat{\underline{V}}(\alpha_1, \gamma_3) = \big(V_1(\alpha_1,\ t\gamma_3), tV_3(\alpha_1, t\gamma_3)\big), \qquad \gamma_3 \in (-\tfrac{1}{2},\ \tfrac{1}{2}) := J.$$

Using the Korn inequality, we get

$$\begin{aligned}
&\int_{\tilde{\Omega}} \left(V_{1,1}^2 + \frac{1}{2}(V_{1,3} + V_{3,1})^2 + V_{3,3}^2\right) d\alpha_1\, d\alpha_3 \\
&\quad = \int_{I\times J} \left(\hat{V}_{1,1}^2 + \frac{1}{2}t^{-2}(\hat{V}_{1,3} + \hat{V}_{3,1})^2 + t^{-4}\hat{V}_{3,3}^2\right) d\alpha_1\, d\gamma_3 \\
&\quad \ge ct\left(|\hat{V}_1|_{1,I\times J}^2 + |\hat{V}_3|_{1,I\times J}^2\right) \\
&\quad = c\int_{\tilde{\Omega}} \left(V_{1,1}^2 + t^2V_{1,3}^2 + t^2V_{3,1}^2 + t^4V_{3,3}^2\right) d\alpha_1\, d\alpha_3,
\end{aligned}$$

where c is independent of $\underline{V}$ and t. Thus,

$$(\text{A.16}) \qquad \|V_{3,1}\|_{L_2(\tilde{\Omega})} \le Ct^{-1/2}\ |||\underline{V}|||_{2D}, \qquad \underline{V} \in \mathcal{U}^{2D}.$$

Proceeding as in (A.6), we have

$$|V_i(1,\alpha_3)|^2 \le \frac{2}{\epsilon}\int_I |V_i(\tau,\alpha_3)|^2\, d\tau + \frac{2\epsilon}{3}\int_I |V_{i,1}(\tau,\alpha_3)|^2\, d\tau.$$

Then, integrating with respect to α_3, recalling (A.15) and (A.16) and substituting $\epsilon = 1$ if $i = 2$ and $\epsilon = t$ if $i = 3$, we obtain

$$\|V_2(1,\cdot)\|_{L_2(\tilde{\Gamma})} \le ckt^{1/2}|||\underline{V}|||_{2D}, \qquad \|V_3(1,\cdot)\|_{L_2(\tilde{\Gamma})} \le ck^2|||\underline{V}|||_{2D}$$

for all $\underline{V} \in \mathcal{U}^{2D}$. The assertion now follows from these inequalities, together with (A.14). □

REFERENCES

[1] D. Arnold and R. Falk, *The boundary layer for the Reissner–Mindlin plate model*, SIAM J. Math. Anal., 21(1990), pp. 281–313.

[2] ———, *Edge effects in the Reissner–Mindlin plate theory*, in Analytical and Computational Models of Shells, Vol. 3, A. K. Noor, T. Belytschko, and J. C. Simo, eds., The American Society of Mechanical Engineers, New York, 1989.

[3] I. Babuška and J. Pitkäranta, *The plate paradox for hard and soft simple support*, SIAM J. Math. Anal., 21 (1990), pp. 551–576.

[4] P. Ciarlet, *Plates and Junctions in Elastic Multi-Structures, An Asymptotic Analysis*, Masson, Paris, 1990.

[5] P. Ciarlet and P. Destuynder, *A justification of the two-dimensional linear plate model*, J. Meć. Théor. Appl., 18 (1979), pp. 315–344.

[6] P. Ciarlet and B. Miara, *Justification of the two-dimensional equations of a linearly elastic shallow shell*, Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie, preprint 1990.

[7] P. Destuynder, *A classification of thin shell theories*, Acta Appl. Math., 4 (1985), pp. 15–63.

[8] ———, *Une théorie asymptotique des plaques minces en élasticité linéaire*, Masson, Paris, New York, 1986.

[9] D. A. Danielson, *Improved error estimates in the linear theory of thin elastic shells*, Proc. Nederl. Akad. Wetensch., 874 (1971), pp. 294–300.

[10] K. O. Friedrichs and R. F. Dressler, *A boundary layer theory for elastic plates*, Comm. Pure Appl. Math., 14 (1961), pp. 1–33.

[11] W. T. Koiter, *On the foundations of the linear theory of thin elastic shells*, Proc. Kon. Nederl. Akad. Wetensch., B 73 (1970), pp. 169–195.

[12] ———, *On the stability of elastic equilibrium, Ch.* 5, *Shells with finite deflections,* Ph.D. thesis, Delft, Paris, Amsterdam, 1945; English translation, NASA, Tech. Report TT F 10, 833, 1967.

[13] D. Morgenstern, *Herleitung der Plattentheorie aus der Dreidimensionalen Elastizitätstheorie*, Arch. Rational Mech. Anal., 4 (1959), pp. 145–152.

[14] J. Nečas and I. Hlavacek, *Mathematical Theory of Elastic and Elasto-Plastic Bodies, An Introduction*, Elsevier, Amsterdam, New York, 1981.

[15] R. P. Nordgren, *A bound on the error in plate theory*, Quart. Appl. Math., 28 (1971), pp. 587–595.

[16] V. V. Novozhilov, *The Theory of Thin Shells*, P. G. Lowe, transl., J. R. M. Radok, ed., Noordhoff–Groningen, the Netherlands, 1959.

[17] J. Pitkäranta, *The problem of membrane locking in finite element analysis of cylindrical shells*, Institute of Mathematics, Helsinki University of Technology, preprint A290, 1991.

[18] É. Sanchez-Palencia, *Statique et dynamique des coques minces.* I. *Cas de flexion pure non inhibée*, C. R. Acad. Sci. Paris, Sér. I, 309 (1989), pp. 411–417.

[19] ———, *Statique et dynamique des coques minces.* II. *Cas de flexion pure inhibée—Approximation membranaire*, C. R. Acad. Sci. Paris, Sér. I, t.309 (1989), pp. 531–537.

[20] J. L. Sanders, *An improved first-approximation theory of thin shells,* NASA, Tech. Report, R-24 (1959).

[21] J. G. Simmonds, *An improved estimate for the error in the classical, linear theory of plate bending*, Quart. Appl. Math., October, (1971), pp. 439–446.

[22] A. N. Tihonov, *Solution of incorrectly formulated problems and the regularization method*, Dokl. Acad. Nauk SSSR, 151 (1963), Soviet Math. Dokl., 4 (1963), pp. 1035–1038.

SIAM J. MATH. ANAL.
Vol. 24, No. 1, pp. 23-35, January 1993

WEAK SOLUTIONS OF SOME QUASILINEAR ELLIPTIC EQUATIONS WITH DATA MEASURES*

NOUR EDDINE ALAA† AND MICHEL PIERRE†

Abstract. Existence and uniqueness of weak solutions for some quasilinear elliptic equations with data measures and arbitrary growth with respect to the gradient are studied. Usual techniques based on a priori L^∞-bounds for the solutions and its gradient do not apply so that a new approach is needed. Various necessary or sufficient conditions are obtained on the data for existence. Relationship between existence of supersolutions and solutions is considered. Finally, sharp uniqueness results for weak solutions are given.

Key words. quasilinear equations, elliptic, convex nonlinearities, data measures, nonlinear capacities

AMS(MOS) subject classifications. 35J65, 35J25, 31C45

1. Introduction. This paper describes some results concerning existence and uniqueness of weak solutions for equations of the form

$$\alpha u - \Delta u = j(x, u, \nabla u) + \lambda f, \quad u \geqq 0 \quad \text{on } \Omega \tag{1.1}$$

$$u = 0 \quad \text{on } \delta\Omega, \tag{1.2}$$

where Ω is a bounded open set in $\mathbb{R}^N$, α, λ are nonnegative real numbers, j: $\Omega \times \mathbb{R} \times \mathbb{R}^N \to [0, \infty)$ measurable and continuous with respect to u and ∇u, and f: $\Omega \to [0, \infty)$ a nonnegative integrable function or, more generally a given finite nonnegative measure on Ω. We are particularly interested in situations where f is irregular and where the growth of j with respect to ∇u is arbitrary and, in particular, larger than $|\nabla u|^2$ for large $|\nabla u|$. The fact that f is not regular requires that one deals with "weak" solutions for which ∇u and even u itself are not bounded. As a consequence the techniques usually used to prove existence and based on a priori L^∞-estimates on u and ∇u fail. Let us make this more precise on a model problem like

$$\begin{gathered} u \in W_0^{1,p}(\Omega), \\ \alpha u - \Delta u = |\nabla u|^p + \lambda f \quad \text{on } \Omega, \end{gathered} \tag{1.3}$$

where $|\cdot|$ denotes the $\mathbb{R}^N$-euclidian norm and $p > 1$.

If $p \leqq 2$, the method of sub- and supersolutions can be used to prove existence in (1.3) if f is regular enough. For instance, if $\alpha > 0$ and $f \in L^\infty(\Omega)$, then (1.3) has a solution since $u_1 \equiv 0$ is a subsolution and $u_2 \equiv \|f\|_\infty/\alpha$ is a supersolution (see, e.g., [10], [4], [3]). The situation is quite different if $p > 2$: for instance a *size condition* is necessary on λf to have existence in (1.3) even if f is very regular. If $\alpha = 0$, the situation is even more different since the critical value is then $p = 1$. Indeed, as proved in next section, existence in (1.3) with $\alpha = 0$ and $p > 1$ requires that f be *small enough* and *regular enough.*

It is proved in Lions [12] that, if (1.3) has a nonnegative supersolution in $W_0^{1,\infty}(\Omega)$, then (1.3) has a solution (no matter the value of p). Note that here the supersolution is required to vanish at the boundary. This provides an a priori pointwise estimate for ∇u at the boundary. The boundedness on ∇u on the whole set Ω is then obtained by a maximum principle applied to the equation satisfied by $|\nabla u|^2$. The convexity of $r \to |r|^p$ is there an essential ingredient.

* Received by the editors April 8, 1992; accepted for publication May 19, 1992.

† Department of Mathematics, BP 239, University of Nancy I-54506-Vandoeuvre-lès-Nancy, France.
This work has been supported by the University of Nancy I, URA CNRS 750 and INRIA-Lorraine, Projet NUMATH.

Obviously this approach fails to provide existence when f is not regular enough to expect $W^{1,\infty}$-solutions and new techniques have to be used. We describe some of them here.

Another difficulty is that uniqueness does not hold in general for weak solutions (although it does for regular solutions): if $f \equiv 0$ and Ω is a ball, (1.3) can have nonzero solutions (see § 3). Moreover, they are not regular: this proves that, even for good f, good a priori estimates can only be obtained for some of the solutions.

We prove in § 3 that uniqueness of solutions of (1.3) hold if and only if $p < N/(N-1)$ and that uniqueness of "strong" (i.e., in $W^{1,\infty}$) solutions hold. Section 2 is devoted to necessary conditions on the data to get existence of weak solutions in (1.1). Section 3 deals with uniqueness. Section 4 presents an existence result based on an isoperimetric inequality, for linear growth and any finite measure f. In the last section, we prove that existence of weak supersolutions implies existence of weak solutions in the case of (sub)quadratic growth. The difficulties in this section are similar to those in [8], and the techniques are of the same spirit.

2. Necessary conditions for existence. Throughout this paper, we denote by Ω a bounded open set in $\mathbb{R}^N$ with regular boundary. In this section, we are given

$$f \text{ a nonnegative finite measure on } \Omega \tag{2.1}$$

and

$$j\colon \Omega\times\mathbb{R}^N \to [0,\infty[\text{ such that}$$

$$j \text{ is measurable, a.e. } x,\ r \to j(x,r) \text{ is convex, continuous,} \tag{2.2}$$

$$\forall r \in \mathbb{R}^N \quad j(\cdot, r) \in L^1(\Omega), \tag{2.3}$$

$$j(x,0) = \min\{j(x,r),\, r\in\mathbb{R}^N\} = 0. \tag{2.4}$$

For $\lambda \in \mathbb{R}$, we consider the problem

$$\begin{aligned} &u \in W_0^{1,1}(\Omega), && j(\cdot, \nabla u) \in L^1_{\mathrm{loc}}(\Omega),\\ &-\Delta u \geqq j(\cdot,\nabla u) + \lambda f && \text{in } \mathscr{D}'(\Omega). \end{aligned} \tag{2.5}$$

We first state that, if $j(\cdot, r)$ is superlinear at infinity, then there exists $\lambda^* < \infty$ such that (2.5) does not have any solution for $\lambda > \lambda^*$. Moreover, f should also be regular enough. A rather sharp superlinearity condition on j is given next where the x-dependence is taken into account. We assume

(2.6) there exists ω open in Ω and $J\colon \mathbb{R}\to[0,\infty)$ convex, continuous with $J(0) = J'(0^+) = 0$

and

$$j(x,r) \geqq J(|r|) \quad \text{a.e. } x \in \omega, \tag{2.7}$$

$$\int^{+\infty} \frac{ds}{J(s)} < \infty, \tag{2.8}$$

$$\int_\omega f > 0. \tag{2.9}$$

THEOREM 2.1. *Assume* (2.1)-(2.4), (2.6)-(2.9) *hold. Then there exists* $\lambda^* < \infty$ *such that* (2.5) *does not have any solution for* $\lambda > \lambda^*$. *Moreover, when* (2.5) *has a solution,* *then for all nonnegative* φ *in* $C_0^\infty(\omega)$,

$$\lambda\int_\omega \varphi f \leqq \int_\omega \varphi J^*\left(\frac{|\nabla\varphi|}{\varphi}\right), \tag{2.10}$$

where J^* *is the convex conjugate function of* J.

Remarks. The growth condition (2.8) is sharp (see remarks in § 3). The introduction of ω allows the growth of $j(x, r)$ in r to depend on x. In particular, $j(x, r)$ can be sublinear for some x. However, it should be superlinear on some part of the support of f. Examples can be easily constructed in dimension 1, showing the necessity of (2.9) (see [2]).

The condition (2.10) is at the same time a size and a regularity condition on f. It is similar to those obtained in [6] for semilinear problems of the form "$-\Delta u = u^p + \lambda f$". We refer to [6], [1] for various discussions on the meaning of (2.10) (in particular for the power case) and for its relationship with nonlinear capacities.

Proof of Theorem 2.1. Assume u is a solution of (2.5). By (2.5)-(2.7), we have

$$-\Delta u \geqq J(|\nabla u|) + \lambda f \quad \text{in } \mathscr{D}'(\omega). \tag{2.11}$$

Let $\varphi \in C_0^\infty(\omega)$, $\varphi \geqq 0$. Multiply (2.11) by φ and integrate to obtain

$$\lambda\int_\omega \varphi f \leqq \int_\omega \nabla u \nabla\varphi - \varphi J(|\nabla u|) \leqq \int_\omega \varphi\left[|\nabla u|\frac{|\nabla\varphi|}{\varphi} - J(|\nabla u|)\right]. \tag{2.12}$$

If we recall that the conjugate function of J is defined by

$$\forall s \in \mathbb{R}^N, \qquad J^*(s) = \sup\{\alpha s - J(\alpha),\ \alpha \in \mathbb{R}\}, \tag{2.13}$$

we see that (2.12) implies

$$\forall\varphi \in C_0^\infty(\omega), \qquad \lambda\int_\omega \varphi f \leqq \int_\omega \varphi J^*\left(\frac{|\nabla\varphi|}{\varphi}\right). \tag{2.14}$$

This proves (2.10).

Let us now prove that this implies that λ is finite (whence the existence of λ^*). By density, (2.12) and (2.14) remain valid for $\varphi \in W_0^{1,\infty}(\omega)$. By (2.14) and (2.9), the existence of λ^* as in Theorem 2.1 will be proved if we can construct φ such that

$$\varphi \in W_0^{1,\infty}(\omega),\ \varphi > 0 \quad \text{on } \omega, \qquad \int_\omega \varphi J^*(|\nabla\varphi|/\varphi) < \infty. \tag{2.15}$$

Without loss of generality, it can be assumed that ω is a ball and, by translation, the ball $B(0, \varepsilon)$. We set

$$F(r) = \int_0^r \frac{ds}{J(s) + M},$$

where $M > 0$ is chosen large enough so that (see (2.8))

$$\eta = F(+\infty) \leqq \varepsilon.$$

We then introduce

$$\chi(r) = F^{-1}(r), \chi\colon [0, \eta) \to [0, \infty),$$

$$\varphi(x) = \begin{cases} \dfrac{1}{J(\chi(|x|)) + M} & 0 \leqq r \leqq \eta, \\ 0 & \eta \leqq r \leqq \varepsilon. \end{cases} \tag{2.16}$$

Writing $\varphi(x) = h(|x|)$, we check that

$$h'(r) = -h(r)J'(\chi(r))$$

$$I = \int_\omega \varphi J^*\left(\frac{|\nabla\varphi|}{\varphi}\right) = \omega_N \int_0^\eta r^{N-1}h(r)J^*(J'(\chi(r)))\,dr.$$

But, $J^*(J'(\chi(r))) = \chi(r)J'(\chi(r)) - J(\chi(r))$ so that

$$I \leqq \omega_N\eta^{N-1}\int_0^\eta -h'\chi - hJ(\chi) = \omega_N\eta^{N-1}\left\{[h\chi]_{r=\eta}^{r=0} + \int_0^\eta h(\chi' - J(\chi))\right\}.$$

Using (2.16), $\lim_{s\to\infty} s/J(s) = 0$, $\chi' = J(\chi) + M$, we deduce

$$I \leqq \omega_N\eta^{N-1}M\int_0^\eta h(r)\,dr < +\infty.$$

This proves (2.15) with $\omega = B(0, \varepsilon)$ and Theorem 2.1.

We now look at the problem (1.3) when $\alpha > 0$. As noticed in the introduction, if f is regular and $1 < p < 2$, then (1.3) has a solution. In other words, no size condition is required on f. However,

- a regularity condition is required on f as soon as $p > 1$,
- if, moreover, $p > 2$, then a size condition is also required.

This is the purpose of the two following results stated for data f, j satisfying (2.1)-(2.4) and

$$\text{(2.17)} \qquad \exists p > 1,\ C_1, C_2 > 0, \qquad j(x, r) \geqq C_1|r|^p - C_2.$$

We consider the following problem when $\alpha \geqq 0$:

$$\text{(2.18)} \qquad u \in W_0^{1,1}(\Omega), \qquad j(\,\cdot\,, \nabla u) \in L^1_{\mathrm{loc}}(\Omega)$$

$$\text{(2.19)} \qquad \alpha u - \Delta u \geqq j(\,\cdot\,, \nabla u) + \lambda f \quad \text{in } \mathscr{D}'(\Omega).$$

PROPOSITION 2.2. *Assume* (2.1)-(2.4), (2.17) *hold. Assume* (2.18), (2.19) *has a solution for some* $\lambda > 0$. *Then the measure* f *does not charge the sets of* $W^{1,p'}$*-capacity zero.*

Remark. We recall that a compact set $K \subset \Omega$ is of $W^{1,p'}$-capacity zero if there exists a sequence of C_0^∞-functions φ_n greater than 1 on K and converging to zero in $W^{1,p'}$. The above statement means that

$$\text{(2.20)} \qquad (K \text{ compact, } W^{1,p'}\text{-capacity } (K) = 0) \Rightarrow \int_K f = 0.$$

Obviously, this is not true for any measure f as soon as $N > p'$ or $p > N/(N-1)$ (see, e.g., [7] and the references there for more details).

PROPOSITION 2.3. *Assume* (2.1)-(2.4) *and* (2.17) *hold with* $p > 2$. *Then there exists* $\lambda^* < \infty$ *such that* (2.18), (2.19) *does not have any solution for* $\lambda > \lambda^*$.

Proof of Proposition 2.2. From (2.19), (2.17), we get the following inequality:

$$\text{(2.21)} \qquad \alpha u - \Delta u \geqq C_1|\nabla u|^p - C_2 + \lambda f.$$

Let K be a compact set of $W^{1,p'}$-capacity zero and φ_n on a sequence of C_0^∞-functions such that

$$\text{(2.22)} \qquad \varphi_n \geqq 1 \quad \text{on } K, \quad \|\varphi_n\|_{W^{1,p'}} \xrightarrow{n\to\infty} 0, \quad 0 \leqq \varphi_n \leqq 1.$$

Multiplying (2.21) by $\chi_n = \varphi_n^{p'}$ leads to

$$\lambda \int_\Omega f\chi_n + C_1 \int_\Omega \chi_n |\nabla u|^p \leqq C_2 \int_\Omega \chi_n + \alpha \int_\Omega u\chi_n + \int_\Omega \nabla u \nabla \chi_n. \tag{2.23}$$

We use $\nabla \chi_n = p' \varphi_n^{p'-1} \nabla \varphi_n$ and Young's inequality to treat the last integral above:

$$\int_\Omega \nabla u \nabla \chi_n \leqq \varepsilon \int_\Omega \chi_n |\nabla u|^p + C_\varepsilon \int_\Omega |\nabla \varphi_n|^{p'}. \tag{2.24}$$

Due to (2.22), passing to the limit in (2.23), (2.24) with ε small enough easily leads to

$$\lambda \int_K f \leqq 0. \tag{2.25}$$

Proof of Proposition 2.3. Let us consider a nonnegative function $\varphi \in W_0^{1,\infty}(\Omega)$ such that there are constants k_1, k_2 with

$$k_1\, d(x, \delta\Omega) \leqq \varphi(x) \leqq k_2\, d(x, \delta\Omega) \quad \forall x \in \Omega. \tag{2.26}$$

Let us also introduce the solution of

$$\begin{aligned} &\theta \in W_0^{1,\infty}(\Omega), \\ &-\Delta\theta = \varphi \quad \text{on } \Omega. \end{aligned} \tag{2.27}$$

Multiply inequality (2.21) by φ and integrate by parts to obtain

$$\begin{aligned} \lambda \int_\Omega f\varphi + C_1 \int_\Omega \varphi |\nabla u|^p &\leqq C_2 \int_\Omega \varphi + \int_\Omega \nabla u(\alpha \nabla\theta + \nabla\varphi) \\ &\leqq C_2 \int_\Omega \varphi + \varepsilon \int_\Omega \varphi |\nabla u|^p + C_2 \int_\Omega \varphi^{1-p'} |\alpha\nabla\theta + \nabla\varphi|^{p'}. \end{aligned}$$

Again we prove, as in Theorem 2.1, existence of λ^* by showing that the last integral is bounded. But due to (2.26), $\varphi^{1-p'}$ is integrable on Ω if $p'-1<1$ or $p>2$. On the other hand, $\nabla\theta$, $\nabla\varphi$ are in $L^\infty(\Omega)$.

Remark. A sharper result can be obtained in the same way for functions j satisfying $j(x, r) \geqq C_1 J(|r|) - C_2$, where J is a convex function such that $\int_\Omega \varphi J^*(C/\varphi) < \infty$ for C large.

3. An existence result for any finite measure.

THEOREM 3.1. *Let j satisfy* (2.2)-(2.4) *and assume there exist C_1 in $L^{N+\eta}(\Omega)$, $\eta > 0$ and C_2 in $L^1(\Omega)$ such that*

$$\text{p.p.}x \quad \forall r \quad j(x, r) \leqq C_1(x)|r| + C_2(x). \tag{3.1}$$

Then for all $\alpha \geqq 0$, all finite measure f and all $\lambda \in \mathbb{R}$, the problem (2.18), (2.19) *has a solution.*

The main ingredient in the proof is the isoperimetric inequality that we use under the following form (see, e.g., [13], [14] and the references there).

LEMMA 3.2. *Let $u \in W_0^{1,1}(\Omega)$. Then*

$$-\frac{d}{dt} \int_{[|u|>t]} |\nabla u| \geqq N\omega_N^{1/N} \mu(t)^{1-1/N}, \tag{3.2}$$

where ω_N is the measure of the unit ball and

$$\mu(t) = \text{measure } \{x \in \Omega : |u(x)| > t\}. \tag{3.3}$$

Proof of Theorem 3.1. We regularize the problem (2.18), (2.19) by taking regular functions f_ε converging to f in the sense of measures, $j_\varepsilon(x,\cdot)$ a regular approximation of $j(x,\cdot)$ satisfying (3.1) uniformly. Then, for all $\varepsilon>0$, there exists u_ε solution of

$$\text{(3.4)}\qquad \begin{aligned} &u_\varepsilon\in W_0^{1,\infty}(\Omega),\\ &(\alpha+\varepsilon)u_\varepsilon-\Delta u_\varepsilon=j_\varepsilon(x,\nabla u_\varepsilon)+\lambda f_\varepsilon \quad \text{in } \Omega. \end{aligned}$$

Indeed, the constant function $\lambda\|f_\varepsilon\|_{L^\infty(\Omega)}/(\alpha+\varepsilon)$ is a supersolution of (3.4), and a subsolution is given by w_ε solution of

$$\text{(3.5)}\qquad \begin{aligned} &w_\varepsilon\in W_0^{1,\infty}(\Omega),\\ &(\alpha+\varepsilon)w_\varepsilon-\Delta w_\varepsilon=\lambda f_\varepsilon \quad \text{in } \Omega, \end{aligned}$$

which satisfies $w_\varepsilon\leqq\lambda\|f_\varepsilon\|_{L^\infty(\Omega)}/(\alpha+\varepsilon)$. Since the growth of j_ε with respect to r is subquadratic, the classical theory applies (see [4]): there exists u_ε solution of (3.4).

We will prove the existence of M independent of ε such that

$$\text{(3.6)}\qquad \int_\Omega|\nabla u_\varepsilon|^q\leqq M,\qquad q=(N+\eta)'.$$

Let us first show how the existence of u solution of (2.18), (2.19) follows from (3.6) by passing to the limit in (3.4). From (3.6), since $u_\varepsilon=0$ on Ω, we first deduce

$$\|u_\varepsilon\|_{L^1}\leqq M_1,\quad \text{independent of } \varepsilon.$$

From (3.6) again and (3.4), (3.1), we obtain $\|\Delta u_\varepsilon\|\leqq M_1$. This yields compactness of u_ε in $W_0^{1,r}(\Omega)$ for all $r<N/(N-1)$ (this can be obtained by duality of the compactness of $h\to v$ from $L^r(\Omega)$ into $L^\infty(\Omega)$ for $r>N$, where v is the solution of

$$v\in W_0^{1,2}(\Omega),\qquad -\Delta v=h_{x_i}\quad\text{on } \Omega.$$

See, e.g., [11], [9]). We then pass to the limit in (3.4) to conclude.

We now drop the ε-dependence in (3.4) to make estimates on ∇u_ε and reach (3.6). For $0<t<t+h$, we introduce the function $z(\cdot)$ defined by

$$\text{(3.7)}\qquad z(r)=\begin{cases} 1 & r\geqq t+h,\\ (r-t)^+/h & t\leqq r\leqq t+h,\\ 0 & 0\leqq r\leqq t,\\ -z(-r) & r<0. \end{cases}$$

We multiply (3.4) by $z(u)$ and integrate by parts to obtain

$$(\alpha+\varepsilon)\int_\Omega uz(u)+\int_\Omega z'(u)|\nabla u|^2=\int_\Omega z(u)(j(\cdot,\nabla u)+\lambda f).$$

Using the growth assumption (3.1) and the definition of z, we deduce

$$\text{(3.8)}\quad \frac{1}{h}\int_{t<|u|<t+h}|\nabla u|^2\leqq\int_{|u|>t}C_1|\nabla u|+C_2+\lambda\|f\|_{L^1(\Omega)}\leqq C_q\left(\int_{|u|>t}|\nabla u|^q\right)^{1/q}+C_\lambda,$$

where $q=(N+\eta)'=N/(N-1)-\varepsilon(\eta)$, $\varepsilon(\eta)>0$, $C_q=\|C_1\|_{L^{N+\eta}}$ and $C_\lambda=\|C_2\|_{L^1(\Omega)}+\lambda\|f\|_{L^1(\Omega)}$. We now assume $N\geqq2$ so that $q<2$ and use that

$$\text{(3.9)}\quad \frac{1}{h}\int_{t<|u|<t+h}|\nabla u|^q\leqq\left(\frac{1}{h}\int_{t<|u|<t+h}|\nabla u|^2\right)^{q/2}\left(\frac{\mu(t)-\mu(t+h)}{h}\right)^{(2-q)/2},$$

$$\text{(3.10)}\quad \frac{1}{h}\int_{t<|u|<t+h}|\nabla u|\leqq\left(\frac{1}{h}\int_{t<|u|<t+h}|\nabla u|^q\right)^{1/q}\left(\frac{\mu(t)-\mu(t+h)}{h}\right)^{(q-1)/q},$$

where $\mu(t) = \text{measure } ([|u| > t])$. We take the qth power of (3.10), multiply by the square of (3.9) to obtain

$$\left(\frac{1}{h}\int_{t<|u|<t+h} |\nabla u|\right)^q \left(\frac{1}{h}\int_{t<|u|<t+h} |\nabla u|^q\right) \leqq \left(\frac{1}{h}\int_{t<|u|<t+h} |\nabla u|^2\right)^q \left(\frac{\mu(t)-\mu(t+h)}{h}\right).$$

We now plug (3.8) into the above inequality, and by letting h go to zero, we obtain a differential inequality for $g(t) = \int_{|u|>t} |\nabla u|^q$, namely,

$$\left(-\frac{d}{dt}\int_{|u|>t} |\nabla u|\right)^q (-g'(t)) \leqq (C_q g(t)^{1/q} + C_\lambda)^q (-\mu'(t)) \leqq (D_q g(t) + D_\lambda)(-\mu'(t)).$$

By the isoperimetric inequality (3.2), this also gives

$$-g'(t) + C_N g(t)\mu'(t)\mu(t)^{-q(1-1/N)} \leqq -\mu'(t)\mu(t)^{-q(1-1/N)}\lambda_N$$

with $C_N = D_q(N\omega_N^{1/N})^{-q}$, $\lambda_N = D_\lambda(N\omega_N^{1/N})^{-q}$. This can be rewritten as

$$-\frac{d}{dt}(e^{-k\mu(t)^\alpha} g(t)) \leqq \frac{d}{dt}[e^{-k\mu(t)^\alpha}]\frac{\lambda_N}{C_N}, \quad \alpha = 1 - q\frac{N-1}{N} > 0, \quad k\alpha = C_N.$$

Integrating from $t = 0$ to $t = \|u\|_\infty$ leads to (using $\mu(\|u\|_\infty) = 0$)

$$e^{-k\mu(0)^\alpha} g(0) \leqq \frac{\lambda_N}{C_N}.$$

Since $\mu(0) \leqq |\Omega|$, this yields

$$\int_\Omega |\nabla u|^q \leqq e^{k|\Omega|^\alpha}\frac{\lambda_N}{C_N},$$

which is (3.6). The adaptation is obvious if $N = 1$.

Remark. It is very likely that (3.1) can be replaced by a more general condition of the form

$$j(x, r) \leqq C_1 J(r) + C_2,$$

where $\int^\infty (ds/J(s)) = \infty$ (see (2.7), (2.8)). But adapting directly the above proof to that case does not seem straightforward.

4. About uniqueness of weak solutions. Although uniqueness and order-preserving hold for *regular* solutions (say, in $W_0^{1,\infty}$) of (2.19), they completely fail for solutions in the weaker sense (2.18), as proved by the following simple examples.

Example 4.1. Let $N > 2$, $\Omega = B(0, 1)$ and $u(x) = -(N-2)\ln|x|$. Then

$$\begin{aligned} &u \in W_0^{1,2}(\Omega), \\ &-\Delta u = |\nabla u|^2 \quad \text{in } \Omega, \end{aligned} \tag{4.1}$$

and $u \equiv 0$ is also solution of (4.1).

Example 4.2. Let $N > 1$, $\Omega = B(0, 1)$, $p > N/(N-1)$, $p \neq 2$, and

$$u(x) = C(1 - |x|^\alpha) \quad \text{with } \alpha = \frac{(p-2)}{(p-1)}, \quad \alpha C = N - 2 + \alpha (>0).$$

Then

$$u \in W_0^{1,p}(\Omega), -\Delta u = |\nabla u|^p \quad \text{in } \Omega. \tag{4.2}$$

Next, we make precise the class of solutions for which order-preserving and uniqueness hold. We consider the following problem where f, j are defined in (2.1)-(2.4) and $\alpha \geqq 0$:

$$u \in W_0^{1,1}(\Omega), \quad j(\cdot, \nabla u) \in L^1(\Omega), \tag{4.3}$$

$$\alpha u - \Delta u = j(\cdot, \nabla u) + f \quad \text{in } \mathscr{D}'(\Omega). \tag{4.4}$$

We call subsolution (respectively, supersolution) of (4.3), (4.4) a function u satisfying (4.3) and (4.4) with "$=$" replaced by "$\leqq$" (respectively, $\geqq$). We will say that u is a *regular* subsolution, supersolution, or solution of (4.3), (4.4) if, moreover,

$$\exists \varepsilon > 0, \quad \exists \vec{a} \in L^{N+\varepsilon}(\Omega, \mathbb{R}^N), \quad \vec{a}(x) \in \delta j(x, \nabla u(x)) \text{ a.e. } x \in \Omega, \tag{4.5}$$

where $\delta j(x, \cdot)$ is the subdifferential of $j(x, \cdot)$.

Theorem 4.3. *Let u be a supersolution of* (4.3), (4.4), *and let $\hat{u}$ be a* regular *subsolution of* (4.3), (4.4) *associated with $\hat{j}, \hat{f}$ satisfying* (2.1)-(2.4) *and $j \geqq \hat{j}, f \geqq \hat{f}$. Then $u \geqq \hat{u}$.*

Corollary 4.4. *Under assumptions* (2.1)-(2.4), *regular solutions of* (4.3), (4.4) *are unique. Moreover, the regular solution is the smallest solution when it exists.*

Corollary 4.5. *Assume* (2.1)-(2.4) *and*

$$\exists p < \frac{N}{(N-1)}, \, C_1, \, C_2 > 0, j(x, r) \leqq C_1 |r|^p + C_2. \tag{4.6}$$

Then solutions of (4.3), (4.4) *are unique.*

Remark. According to Examples 4.1, 4.2, the condition $p < N/(N-1)$ is sharp for the uniqueness of (weak) solutions of (4.3), (4.4). So is the assumption on $\vec{a}$ in the next lemma.

Lemma 4.6. *Let $\vec{a} \in L^{N+\varepsilon}(\Omega, \mathbb{R}^N)$, $\varepsilon > 0$, and w a solution of*

$$w \in W_0^{1,1}(\Omega), \qquad \Delta w \in L^1(\Omega), \tag{4.7}$$

$$\alpha w - \Delta w \leqq \vec{a} \cdot \nabla w \quad \textit{in } \mathscr{D}'(\Omega). \tag{4.8}$$

Then $w \leqq 0$.

Remark. Similar results can be found in [11] for $w \in W^{1,2}(\Omega)$ in a quite more general framework. However, the weaker assumption $w \in W^{1,1}$ makes a significant difference. Note that $w \in W_0^{1,1}$ and $\Delta w \in L^1$ imply that $w \in W_0^{1,p}$ for all $p \in [1, N/(N-1)[$ (see, e.g., [9]). It follows that $\vec{a} \cdot \nabla w$ is in $L^{1+\eta}$ for some $\eta > 0$.

Proof of Lemma 4.6. We give here a direct proof using the same isoperimetric inequality as in § 3. First, by elementary regularization argument, we prove $w^+ \in W_0^{1,1}(\Omega)$ and

$$\alpha w^+ - \Delta(w^+) \leqq \vec{a} \cdot \nabla(w^+) = \vec{a} \cdot \nabla w \cdot 1_{[w>0]} \quad \text{in } \mathscr{D}'(\Omega).$$

In other words, $w \geqq 0$ can be assumed in Lemma 4.6, without loss of generality.

Let f_n be a regular smooth approximation in $L^1(\Omega)$ of $\alpha w - \Delta w$. By classical results (see, e.g., [9]), the solution of

$$\begin{aligned} &w_n \in W_0^{1,\infty}(\Omega) \cap C^2(\Omega), \\ &\alpha w_n - \Delta w_n = f_n \end{aligned} \tag{4.9}$$

converges in $W_0^{1,p}$ to w for all $p \in [1, N/(N-1)]$ (note that this requires uniqueness for the Dirichlet problem in $W_0^{1,1}(\Omega)$). Multiplying (4.9) by $z(w_n)$, where z is defined as in (3.7) leads to

$$\frac{1}{h} \int_{t < w_n < t+h} |\nabla w_n|^2 \leqq \int z(w_n) f_n.$$

Passing to the limit, we obtain that $\nabla w \in L^2([t<w<t+h])$ and

$$(4.10)\qquad \frac{1}{h}\int_{t<w<t+h}|\nabla w|^2 \leqq \int_{w>t} z(w)(\alpha w-\Delta w) \leqq \int_{w>t}|\vec{a}||\nabla w| \leqq C_p\left(\int_{w>t}|\nabla w|\right)^{1/p},$$

where $p=(N+\varepsilon)'=N/(N-1)-\eta(\varepsilon)$, $\eta(\varepsilon)>0$ and $C_p=\|\vec{a}\|_{L^{N+\varepsilon}}$. We now assume $N \geqq 2$ and argue exactly like in the proof of Theorem 3.1 (see (3.9), (3.10)). We use

$$(4.11)\qquad \frac{1}{h}\int_{t<w<t+h}|\nabla w|^p \leqq \left(\frac{1}{h}\int_{t<w<t+h}|\nabla w|^2\right)^{p/2}\left(\frac{\mu(t)-\mu(t+h)}{h}\right)^{(2-p)/2}$$

$$(4.12)\qquad \frac{1}{h}\int_{t<w<t+h}|\nabla w| \leqq \left(\frac{1}{h}\int_{t<w<t+h}|\nabla w|^p\right)^{1/p}\left(\frac{\mu(t)-\mu(t+h)}{h}\right)^{(p-1)/p},$$

where $\mu(t)=$ measure $([w>t])$. We take the pth power of (4.12), multiply by the square of (4.11) and use (4.10) to obtain

$$\left(\frac{1}{h}\int_{t<w<t+h}|\nabla w|\right)^p\left(\frac{1}{h}\int_{t<w<t+h}|\nabla w|^p\right) \leqq C_p^p\int_{w>t}|\nabla w|^p\left(\frac{\mu(t)-\mu(t+h)}{h}\right).$$

By letting h go to zero, we obtain a differential inequality for $g(t)=\int_{w>t}|\nabla w|^p$, namely,

$$(4.13)\qquad \left(-\frac{d}{dt}\int_{w>t}|\nabla w|\right)^p(-g'(t)) \leqq C_p^p g(t)(-\mu'(t)).$$

By the isoperimetric inequality (3.2), this writes

$$-k_p g'(t)+C_p^p g(t)\mu'(t)\mu(t)^{-p(1-1/N)} \leqq 0, \qquad k_p=(N\omega_N^{1/N})^p$$

or

$$-\frac{d}{dt}(e^{-k\mu(t)^\alpha}g(t)) \leqq 0, \quad \alpha=1-p\frac{N-1}{N}>0, \quad k\alpha=C_p^p/k_p.$$

Since $\mu(\infty)=g(\infty)=0$, this implies $g(0)=0$, that is, $\nabla w=0$ and $w=0$ since $w \in W_0^{1,p}(\Omega)$. The adaptation is obvious if $N=1$.

Proof of Theorem 4.3. By difference between (4.3), (4.4) applied to u and $\hat{u}$, we have

$$(4.14)\qquad \alpha(\hat{u}-u)-\Delta(\hat{u}-u) \leqq \hat{j}(\cdot,\nabla\hat{u})-j(\cdot,\nabla u).$$

Since $\hat{u}$ is a regular solution and by the convexity of $j(x,\cdot)$, there exists $\hat{a} \in L^{N+\varepsilon}(\Omega,\mathbb{R}^N)$ such that

$$(4.15)\qquad \hat{j}(\cdot,\nabla u)-\hat{j}(\cdot,\nabla\hat{u}) \geqq \vec{a}\cdot\nabla(u-\hat{u}).$$

Since $j(\cdot,\nabla u) \geqq \hat{j}(\cdot,\nabla u)$, we obtain from (4.14), (4.15),

$$\alpha(\hat{u}-u)-\nabla(\hat{u}-u) \leqq \vec{a}\nabla(\hat{u}-u).$$

Therefore, $w=\hat{u}-u$ satisfies the assumptions of Lemma 4.6. The conclusion follows.

Proof of Corollary 4.4. It is a direct consequence of Theorem 4.3 and definition (4.5) of regular solutions.

Proof of Corollary 4.5. It is sufficient to prove that any solution is regular in the sense of (4.5). Since, for all $r, s \in \mathbb{R}^N$,

$$j(x,r+s)-j(x,r) \geqq \delta j(x,r)\cdot s,$$

we deduce from (4.6) that

$$\delta j(x,r)\cdot s \leqq C_1|r+s|^p+C_2$$

and

$$(4.16)\qquad |\delta j(x,r)| \leqq \hat{C}_1|r|^{p-1}+\hat{C}_2 \quad \text{for } |r|>1.$$

Now, if u is solution of (4.3), (4.4), then $\nabla u \in L^q(\Omega)$ for all $q \in [1, N/(N-1)[$. But from (4.16),

$$|\delta j(x, \nabla u)|^{N+\varepsilon} \leqq C_1' |\nabla u|^{(p-1)(N+\varepsilon)} + C_2'.$$

Since $p < N/(N-1)$, $(p-1)(N+\varepsilon) < N/(N-1)$ for ε small enough. This yields (4.5).

5. An existence result for (sub)quadratic growth. We consider here the more general problem

$$\text{(5.1)} \qquad u \in W_0^{1,1}(\Omega), \qquad j(\cdot, u, \nabla u) \in L^1_{\text{loc}}(\Omega),$$

$$\text{(5.2)} \qquad -\Delta u = j(\cdot, u, \nabla u) + f \quad \text{in } \mathscr{D}'(\Omega),$$

where f is as in (2.1) and j: $\Omega \times \mathbb{R} \times \mathbb{R}^N \to [0, \infty)$ satisfies

$$\text{(5.3)} \qquad j \text{ is measurable, a.e. } x, \ (s, r) \to j(x, s, r) \text{ is continuous,}$$

$$\text{(5.4)} \qquad j \text{ is nondecreasing in } s \text{ and convex in } r,$$

$$\text{(5.5)} \qquad j(x, s, 0) = \min \{ j(x, s, r), r \in \mathbb{R}^N \} = 0.$$

As proved in § 2, if j is superlinear in r, existence does not always hold since a size condition is required on f. A natural question is whether the existence of a nonnegative supersolution implies existence of a solution. This is well known in the "regular" case when moreover the growth of j is at most quadratic. However the proof relies on a priori $W^{1,\infty}$-estimates for the solution. This cannot be expected when f is only a measure. Therefore, a quite different technique has to be used. This is what we do next when j is at most quadratic. The same question remains open for superquadratic growth.

We will assume

$$\text{(5.6)} \qquad j(x, s, r) \leqq C_1(|s|)(|r|^2 + 1)$$

$$\text{(5.7)} \qquad C_1 : [0, \infty) \to [0, \infty) \quad \text{is nondecreasing.}$$

THEOREM 5.1. *Assume* (2.1), (5.3)-(5.7). *Assume there exists w in* $W_0^{1,2}(\Omega)$ *such that*

$$\text{(5.8)} \qquad -\Delta w \geqq j(\cdot, w, \nabla w) + f \quad \text{in } \mathscr{D}'(\Omega).$$

Then, there exists u solution of

$$\text{(5.9)} \qquad \begin{aligned} & u \in W_0^{1,2}(\Omega) \\ & -\Delta u = j(\cdot, u, \nabla u) + \lambda f \quad \text{in } \mathscr{D}'(\Omega) \end{aligned}$$

for all $\lambda \in [0, 1]$.

For the proof of Theorem 5.1, we introduce regularized versions of (5.9) with smaller j_n with linear growth. The corresponding solutions u_n—which exist by § 3—are such that

$$0 \leqq u_n \leqq u_{n+1} \leqq w.$$

For this, we use the monotonicity result of the previous section. The convergence of u_n to a solution of (5.9) will then follow from the next two lemmas.

LEMMA 5.2. *Let* $u \in W_0^{1,1}(\Omega)$, $v \in W_0^{1,2}(\Omega)$ *such that*

$$\text{(5.10)} \qquad 0 \leqq u \leqq v \quad \text{in } \Omega, \quad 0 \leqq -\Delta u.$$

Then, $u \in W_0^{1,2}(\Omega)$ *and*

$$\int_\Omega |\nabla u|^2 \leqq \int_\Omega |\nabla v|^2. \tag{5.11}$$

Remarks. If $v \in W_{\text{loc}}^{1,2}(\Omega)$ only, (5.11) can be replaced by local estimates.

Obviously Lemma 5.2 will provide $W^{1,2}$-estimates for the approximate solution u_n. But, we need a strong convergence in $W^{1,2}$ to pass to the limit in the nonlinear terms. It will be a consequence of the following lemma.

LEMMA 5.3. *Let* $u_n \in W_0^{1,2}(\Omega)$ *converging weakly in* $W_0^{1,2}(\Omega)$ *to* u *and such that*

$$0 \leqq u_n \leqq u, \; -\Delta u_n \geqq 0 \quad \textit{in } \mathscr{D}'(\Omega). \tag{5.12}$$

Then u_n *converges to* u *strongly in* $W_0^{1,2}(\Omega)$.

Proof of Lemma 5.2. Let u_n be a sequence of regular functions converging to u and such that $-\Delta u_n \geqq 0$. We write

$$\int_\Omega |\nabla u_n|^2 = \int_\Omega u_n(-\Delta u_n) \leqq \int_\Omega v(-\Delta u_n) = \int_\Omega \nabla v \nabla u_n \leqq \left(\int_\Omega |\nabla v|^2\right)^{1/2} \left(\int_\Omega |\nabla u_n|^2\right)^{1/2}.$$

The result follows.

Proof of Lemma 5.3. We write

$$\int_\Omega |\nabla(u-u_n)^2 = \int_\Omega \nabla u \nabla(u-u_n) - \int_\Omega \nabla u_n \nabla u + \int_\Omega |\nabla u_n|^2,$$

and we bound the last integral as follows:

$$\int_\Omega |\nabla u_n|^2 = \int_\Omega u_n(-\Delta u_n) \leqq \int_\Omega u(-\Delta u_n).$$

We now use the $W^{1,2}$-weak convergence of u_n towards u to conclude.

Proof of Theorem 5.1. For $n \geqq 1$, we introduce $\hat{j}_n(x, s, \cdot)$ the Yosida-approximation of $j(x, s, \cdot)$, which increases pointwise to $j(x, s, \cdot)$ as n tends to ∞ and satisfies

$$\hat{j}_n \leqq j, \qquad \|\hat{j}_{nt}(x, s, \cdot)\| \leqq n \tag{5.13}$$

where $\hat{j}_{nr}$ denotes a section of the subdifferential of $\hat{j}_n$ with respect to r. Then we set

$$j_n(x, s, r) = \hat{j}_n(x, s, r) \cdot 1_{[w<n]}, \tag{5.14}$$

where w is defined in (5.8). We check that j_n satisfies (5.3), (5.5), is convex in r, and

$$j_n \leqq j \cdot 1_{[w<n]}, \qquad j_n \leqq j_{n+1}. \tag{5.15}$$

We consider the sequence defined by

$$u_1 \in W_0^{1,1}(\Omega), -\Delta u_1 = \lambda f \cdot 1_{[w<1]}, \tag{5.16}$$

$$u_{n+1} \in W_0^{1,1}(\Omega), \; -\Delta u_{n+1} = j_n(\cdot, u_n, \nabla u_{n+1}) + \lambda f \cdot 1_{[w<n]}. \tag{5.17}$$

Let us prove by induction that

$$u_n \leqq \min(w, n) \quad \text{for all } n. \tag{5.18}$$

Indeed, if $w_n = \min(w, n)$, from (5.8) we easily deduce that

$$-\Delta w_n \geqq j(\cdot, w_n, \nabla w_n) \cdot 1_{[w<n]} + \lambda f 1_{[w<n]}. \tag{5.19}$$

For $n=1$, $-\Delta(w_1-u_1)\geqq 0 \Rightarrow u_1 \leqq w_1$. Let us assume $u_n \leqq w_n$. Then, from (5.19) and the monotonicity of j in s, we have

$$-\Delta w_n \geqq j(\cdot, u_n, \nabla w_n)\cdot 1_{[w<n]} + \lambda f \cdot 1_{[w<n]},$$

which by (5.13) implies

$$-\Delta w_n \geqq j_n(\cdot, u_n, \nabla w_n) + \lambda f \cdot 1_{[w<n]}. \tag{5.20}$$

Now, the monotonicity result of Theorem 4.3 can be applied with the function $J:(x,r)\to j_n(x,u_n(x),r)$, which satisfies (2.2)-(2.4). For this, we have to check that u_{n+1} is a regular solution of (5.17), that is, (see (4.3), (4.5))

$$J(\cdot, \nabla u_{n+1}) \in L^1(\Omega), \tag{5.21}$$

$$J_r(\cdot, \nabla u_{n+1}) \in L^{N+\varepsilon}(\Omega, \mathbb{R}^N). \tag{5.22}$$

Since J is linear at infinity by (5.13), (5.21) comes from $u_n \in W^{1,1}(\Omega)$. Now by definition of j_n,

$$\|J_r(\cdot, \nabla u_{n+1})\| \leqq C_n,$$

whence (5.22). Therefore, Theorem 4.3 applies and we deduce $w_n \geqq u_{n+1}$, which proves (5.18) by induction.

By Lemma 5.2 and estimate (5.18), u_n is bounded in $W_0^{1,2}(\Omega)$. Therefore, there exists u in $W_0^{1,2}(\Omega)$ such that (up to a subsequence)

$$u_n \to u \quad \text{strongly in } L^2, \quad \text{a.e. and weakly in } W_0^{1,2}. \tag{5.23}$$

By Lemma 5.3, the convergence holds strongly in $W_0^{1,2}$.

We will now pass to the limit in (5.17). We only have to prove that

$$j_n(\cdot, u_n, \nabla u_{n+1}) \xrightarrow{L^1(\Omega)} j(\cdot, u, \nabla u). \tag{5.24}$$

The convergence holds almost everywhere in (5.24) (up to a subsequence) by construction of j_n and by strong convergence of u_n and ∇u_n in $L^2(\Omega)$. Since

$$j_n(\cdot, u_n, \nabla u_{n+1}) \leqq C_1(|u_n|)(|\nabla u_{n+1}|^2+1)$$

for all $t>0$,

$$\int_{u_n<t} |j_n(\cdot, u_n, \nabla u_{n+1}) - j(\cdot, u, \nabla u)| \xrightarrow[n\to\infty]{} 0. \tag{5.25}$$

On the other hand, multiplying (5.17) by u_n leads to

$$\int_\Omega u_n j_n(\cdot, u_n, \nabla u_{n+1}) \leqq \int_\Omega \nabla u_n \nabla u_{n+1} \leqq C \quad \text{(independent of } n\text{)}.$$

Therefore, for $\varepsilon>0$ there exists t_ε (e.g., $t_\varepsilon = C/\varepsilon$) such that

$$\int_{u_n>t_\varepsilon} |j_n(\cdot, u_n, \nabla u_{n+1})| \leqq \varepsilon \quad \forall n. \tag{5.26}$$

By Fatou's lemma, we also have

$$\int_{u>t_\varepsilon} |j(\cdot, u, \nabla u)| \leqq \varepsilon. \tag{5.27}$$

But (5.25)-(5.27) imply (5.24). This finishes the proof.

Remark. A natural question is now the following. Let $p > 2$. Assume there exists a solution of

$$w \in W_0^{1,p}(\Omega), \qquad -\Delta w \geqq |\nabla w|^p + f. \tag{5.28}$$

Then, does there exist u solution of

$$u \in W_0^{1,p}(\Omega), \qquad -\Delta u = |\nabla u|^p + f? \tag{5.29}$$

Essentially, the approach above applies except for Lemma 5.3: we do not know whether it extends to $p > 2$. Therefore, strong convergence of u_n in $W^{1,p}$ is lacking. The limit u that is obtained is only a supersolution of the problem. The existence of such a supersolution is actually a very general fact, as proved in [5].

Remark. We refer to P. L. Lions [*Journal d'Analyse Mathématique*, 45 (1985), pp. 234-254] for results related to those in § 2.

REFERENCES

[1] D. R. Adams and M. Pierre, *Capacitary strong type estimates in semi-linear problems*, Ann. Inst. Fourier (Grenoble), 41 (1991), pp. 117-135.

[2] N. Alaa, *Etude d'équations elliptiques non linéaires à dépendance convexe en le gradient et à données mesures*, Thèse, Université de Nancy I, Nancy, France, 1989.

[3] H. Amann, *Fixed point equations and nonlinear eigenvalue problems in order Banach Spaces*, SIAM Rev., 18 (1976), pp. 660-709.

[4] H. Amann and M. G. Crandall, *On some existence theorems for semilinear equations*, Indiana Univ. Math. J., 27 (1978), pp. 779-790.

[5] P. Baras, *Semilinear problem with convex nonlinearity*, in Recent Advances in Nonlinear Elliptic and Parabolic Problems, Proc. Nancy 88, Ph. Bénilan, M. Chipot, L. C. Evans, and M. Pierre, eds., Pitman Res. Notes Math. Ser., 1989.

[6] P. Baras and M. Pierre, *Critère d'existence de solutions positives pour des équations semi-linéaires non monotones*, Ann. Inst. H. Poincaré, 2 (1985), pp. 185-212.

[7] P. Baras and M. Pierre, *Singularités éliminables pour des équations semilinéaires*, Ann. Inst. Fourier (Grenoble), 24 (1984), pp. 185-206.

[8] L. Boccardo and T. Gallouet, *Strongly nonlinear elliptic equations having natural growth terms and L^1 data*, Rapport Instituto per le Appl. d. Calc, Univ. di Roma, 1990; Nonlinear Anal. TMA, to appear.

[9] H. Brezis and W. A. Strauss, *Semi-linear second order elliptic equations in L^1*, J. Math. Soc. Japan, 25 (1973), pp. 565-590.

[10] Y. Choquet-Bruhat and J. Leray, *Sur le problème de Dirichlet quasilinéaire d'ordre deux*, C.R. Acad. Sci. Paris, Ser. A, 274 (1972), pp. 81-85.

[11] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, New York, 1977.

[12] P. L. Lions, *Résolution de problèmes elliptiques quasilinéaires*, Arch. Rational Mech. Anal., 74 (1980), pp. 335-353.

[13] J. Mossino, *Inégalités isopérimétriques et applications en physique*, Hermann, Paris, 1984.

[14] J. M. Rakotoson, *Existence of bounded solutions of some quasilinear degenerate equations*, Comm. Partial Differential Equations, 12 (1987), pp. 633-676.

SIAM J. MATH. ANAL.
Vol. 24, No. 1, pp. 36-45, January 1993

003

EXISTENCE OF GLOBAL WEAK SOLUTIONS TO THE DYNAMICAL PROBLEM FOR A THREE-DIMENSIONAL ELASTIC BODY WITH SINGULAR MEMORY*

HAMID BELLOUT†, FREDERICK BLOOM†‡, AND JINDRICH NECAS§

Abstract. A three-dimensional elastic body with memory is considered, for which the memory term is generated by a singular but integrable kernel; the existence of a global weak solution to an associated initial-boundary value problem is established by constructing Galerkin approximations and deriving suitable energy estimates.

Key words. singular viscoelasticity, energy estimates

AMS(MOS) subject classifications. 35Q99, 45K05, 73F15

1. Introduction. In spite of intense efforts by many outstanding mathematicians, most questions concerning the global existence of both weak and classical solutions to the mixed problem, with large data, for the nonlinear wave equation governing the evolution of the displacement vector **u** in an elastic body, i.e.,

$$\rho\frac{\partial^2 u_i}{\partial t^2}-\frac{\partial}{\partial x_j}\left(\frac{\partial W}{\partial e_{ij}}\right)=\rho f_i, \qquad i=1,2,3, \tag{1.1}$$

remain open. Equations (1.1) are set in an open domain $\Omega\subseteq R^n$, $n\geqq 2$, with $e_{ij}=\frac{1}{2}((\partial u_i/\partial x_j)+(\partial u_j/\partial x_i))$ the components of the infinitesimal strain tensor. The evolution equation (1.1) reflects only the presence of an instantaneous elastic response governed by the stress tensor that applies in the case of small strains, namely,

$$\sigma_{ij}^I=\frac{\partial W}{\partial e_{ij}}, \tag{1.2}$$

where $W=W(e_{ij})$, the stored energy function, is supposed (as it will be in this paper) to be twice continuously differentiable, with bounded second derivatives, and positive definite, i.e., for some $\alpha>0$ and all symmetric $n\times n$ matrices,

$$\frac{\partial^2 W}{\partial e_{ij}\partial e_{kl}}\xi_{ij}\xi_{kl}\geqq \hat{\alpha}\|\xi\|^2. \tag{1.3}$$

We also suppose that $W(0)=0$ and $(\partial W/\partial e_{ij})(0)=0$.

For dynamic nonlinear elasticity problems of the type (1.1) in one-space dimension it is well known, e.g., Lax [16], Klainerman and Majda [17], that solutions develop singularities in finite time even for data that are small and compactly supported; existence of weak solutions for such problems follows from the recent work of DiPerna [18] and is based on compensated compactness arguments. For spatial dimension $n\geqq 2$ there are many partial results for the dynamic problem, e.g., local (in time) existence of classical solutions [19] and formation of singularities in finite time; excellent surveys of progress to date on problems involving but small and finite strain may be found, for example, in the expository paper [20] and the text [21].

* Received by the editors May 7, 1990; accepted for publication (in revised form) March 30, 1992.

† Department of Mathematical Sciences, Northern Illinois University, DeKalb, Illinois 60115.

‡ The research of this author was partially supported by National Science Foundation grant DMS-8823037.

§ Mathematical Institute, Charles University, Prague, Czechoslovakia. This author was on leave during Fall, 1989, at the Department of Mathematical Sciences, Northern Illinois University, DeKalb, Illinois 60115.

In this paper we consider an elastic material with memory for which the Cauchy stress has the form

$$\sigma_{ij}=\sigma^I_{ij}+\sigma^M_{ij}. \tag{1.4}$$

In (1.4), σ^I_{ij} is still given by (1.2), while

$$\sigma^M_{ij}=-\int_0^t h(t-\tau)\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(\tau)-\mathbf{e}(t))\,d\tau, \tag{1.5}$$

with $V=V(\mathbf{e})$ an energy function possessing bounded second derivatives and positive definite, i.e., for some $\beta>0$

$$\frac{\partial^2 V}{\partial e_{ij}\partial e_{kl}}\xi_{ij}\xi_{kl}\geqq\beta\|\boldsymbol{\xi}\|^2 \tag{1.6}$$

for all symmetric $n\times n$ matrices $\boldsymbol{\xi}$; as was the case with the stored energy function W, we also assume that $V(0)=0$ and $(\partial V/\partial e_{ij})(0)=0$. The kernel $h(t)$ in (1.5) is singular but integrable and assumed to have (a nonlinear version of) a form due to Boltzman, namely, $h(t)=\lambda h_0(t)$ with $h_0(t)=t^{-\nu}a(t)+r(t), 0<\nu<\frac{1}{2}$, and $\lambda>0$. Both $a(t)$ and $r(t)$ are sufficiently smooth, decreasing, nonnegative functions on $[0,\infty)$ with $a(t)>0$ in some neighborhood of the origin. We suppose, additionally, that $\int_0^\infty h(\tau)\,d\tau<\infty$. Moreover we take the density $\rho\equiv 1$, which opens the door to the possibility of using a Galerkin approximation scheme, and we employ throughout the Langrangian description of the motion relative to the Cartesian coordinates x_i.

Elastic materials with singular memories of the type described by (1.5) have been considered in several recent pieces of work, e.g., Londen [1], Renardy [2], Hrusa and Renardy [3], Hannsgen and Wheeler [4], Hrusa and Renardy [5], Hrusa and Renardy [6], and Renardy [7]. In [1] the author establishes the existence of weak solutions to an abstract integrodifferential equation; the result applies to the one-dimensional viscoelastic model equation

$$u_{tt}=\phi(u_x)_x+\int_{-\infty}^t a'(t-\tau)\psi(u_x(x,\tau))_x\,d\tau+f(x,t), \tag{1.7}$$

provided $\phi=\psi$ and the memory function has a singularity that is stronger than logarithmic. In [2] it is shown that certain singular kernels do not permit the propagation of singularities and have a smoothing effect; these ideas were further expounded upon in [3], all within a one-dimensional framework. The authors in [4] consider the constant coefficient linear problem on a bounded domain and show that the evolution operator is compact for positive time if and only if the kernel of the integral operator is singular; results such as these, of course, point towards the fact that the models with singular kernels should have nicer existence properties than those with regular kernels. In [5] local and global existence in time (for small data) is established for the history-value problem associated with (1.7) when $a'(t)\sim -t^{-\alpha}, 0<\alpha<1$, as $t\to 0^+$; the equation (1.7) is approximated by equations with regular kernels, and energy estimates are used to prove convergence of the corresponding approximate solutions. The work in [6] differs from that in [5] in that $a(t)$ is allowed to have an integrable singularity at $t=0$, and local (but not global) existence of regular solutions for small data is established. In [7] the author provides an alternative existence proof based on coercive properties of the linearized problem; global existence for small data and local existence for large data (of smooth solutions) is proved for $a(t)$ having a singularity at $t=0$ at least as strong as a negative power of t. We note, in passing, that viscoelastic response, of the Boltzmann variety, governed by kernels that are singular at zero, has been noted by several investigators in the area of polymer physics, e.g., Doi and Edwards [8], Rouse [9], Zimm [10], and Laun [11].

The method that we use in the work reported here is very close in spirit to that employed recently by Milota, Nečas, and Šverák [12], where the scalar equation on $\Omega \subset R^n$,

$$u_{tt} - \frac{\partial}{\partial x_i}(a_i(\nabla u)) + \int_0^t h(t-\tau)\Delta u(\mathbf{x}, \tau)\, d\tau = f, \tag{1.8}$$

is studied, with $h(t) = e^{-\alpha t}t^{-\nu}$, $\alpha > 0$, $0 < \nu < 1$. After the work was completed we received a copy of a preprint of [13], where a model viscoelastic problem similar to the one considered here is treated for a scalar displacement u (as opposed to our vector $\mathbf{u}$). In [13] analogous results to those obtained here are produced by use of an entirely different approximation scheme, one that involves adding the term $\varepsilon\Delta u$ to the left-hand side of (2.2) and studying the limit of u^ε as $\varepsilon \to 0$ (i.e., by use of the viscosity method).

We note here that the work in [13] is somewhat flawed: from Lemma A.2 in [13], and the estimate $\|\mathbf{w}^n - \mathbf{v}^n\|_* \leqq \sqrt{\varepsilon_n}\, C$, therein, it follows only that $\|J_s\mathbf{v}^n - \mathbf{v}^n\|_* \leqq \omega(s) + \sqrt{\varepsilon_n}\, \tilde{C}$ for some $\tilde{C} > 0$ and not that $\|J_s\mathbf{v}^n - \mathbf{v}^n\|_* \leqq \omega(s)$, uniformly in n, for some ω: $[0, 1] \to [0, \infty)$, with $\omega(s) \downarrow 0$ as $\sigma \downarrow 0$, as stated by the author in equation (3.26) of [13]. The lemma following (3.27) is then correct, but vacuous, as one of the hypotheses, i.e., (3.26) of [13], is not satisfied. If (3.26) in [13] is replaced by the somewhat weaker statement alluded to above, a lemma corresponding to that which follows (3.27) in [13] may be stated in a manner sufficiently strong, so as to salvage the rest of the argument in § 3 of [13].

Remarks.

1. It is well known (e.g., Coleman [14]) that, for a viscoelastic body of the type under consideration here, the following condition must be satisfied: in an isothermal process starting from equilibrium, the integral of the stress power around a closed path in strain space must be nonnegative. In the present situation this condition is equivalent to the statement that

$$\int_{t_0}^{t_1} \sigma_{ij}\dot{e}_{ij}\, dt \geqq 0 \quad \forall e_{ij} \in C^1([t_0, t_1]), \quad e_{ij}(x, t_0) = e_{ij}(x, t_1), \tag{1.9}$$

where $x \in \Omega$. For such a closed strain path, of course, $\int_{t_0}^{t_1}(\partial W/\partial e_{ij})(\mathbf{e}(t))\dot{e}_{ij}(t)\, dt = W(\mathbf{e}(t))|_{t_0}^{t_1} = 0$; if we take a closed strain path $e_{ij} \in C^1([t_0, t_1])$ and extend it to the interval $[0, t_0)$ by setting $e_{ij}(x, t) = e_{ij}(x, t_0)$, for $0 \leqq t < t_0$, $x \in \Omega$, then we may compute that

$$\begin{aligned}
&\int_{t_0}^{t_1} \dot{e}_{ij}(t)\left(\int_0^t h(t-\tau)\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(\tau) - \mathbf{e}(t))\, d\tau\right) dt \\
&\quad = \int_0^{t_1} \dot{e}_{ij}(t)\left(\int_0^t h(t-\tau)\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(\tau) - \mathbf{e}(t))\, d\tau\right) dt \\
&\quad = \int_0^{t_1}\left(\int_0^t h(t-\tau)\frac{\partial}{\partial t} V(\mathbf{e}(\tau) - \mathbf{e}(t))\, d\tau\right) dt \\
&\quad = \int_0^{t_1} d\tau \int_\tau^{t_1} h(t-\tau)\frac{\partial}{\partial t} V(\mathbf{e}(\tau) - \mathbf{e}(t))\, dt \\
&\quad = \int_0^{t_1} h(t_1 - \tau) V(\mathbf{e}(\tau) - \mathbf{e}(t))\, d\tau \\
&\qquad - \int_0^{t_1} d\tau \int_\tau^{t_1} \frac{\partial h}{\partial t}(t-\tau) V(\mathbf{e}(\tau) - \mathbf{e}(t))\, dt \geqq 0,
\end{aligned} \tag{1.10}$$

and (1.9) is, therefore, satisfied.

2. We emphasize here that the operator $(\partial/\partial x_j)\sigma_{ij}^M$ is continuous in the same topology, i.e., in $L^2((0, T); W^{1,2}(\Omega; R^n))$ as $(\partial/\partial x_j)\sigma_{ij}^I$ and is even compact in time; this result will be established in the analysis that follows. A key point to be noted is that the memory part of the stress tensor σ_{ij}^M allows us to establish an estimate of the form

$$\|\mathbf{u}\|_{W^{\nu/2,2}((0,T);W^{1,2}(\Omega;R^n))} \leqq C(T), \tag{1.11}$$

for any $T > 0$, which, together with various energy estimates, and an interpolation lemma, yields an estimate of the type

$$\|\mathbf{u}\|_{W^{(1+\nu/2)\alpha,2}((0,T);W^{\alpha,2}(\Omega;R^n))} \leqq C'(T) \tag{1.12}$$

for $1/(1+\nu/2) < \alpha < 1$. The estimates (1.11), (1.12) are key ingredients in our global existence result.

2. Galerkin approximations and basic estimates. We assume in all that follows that Ω is a bounded domain, with Lipschitz continuous boundary $\partial\Omega$, and begin by choosing a basis $\{\mathbf{w}^i\}_{i=1}^{\infty}$ in the space $W_0^{1,2}(\Omega; R^n)$; a natural choice for such a basis would consist of the set of eigenfunctions for the problem

$$\frac{\partial}{\partial x_j}\left(\frac{\partial^2 W}{\partial e_{ij}\partial e_{kl}}(0)\frac{\partial u_k}{\partial x_l}\right) + \lambda u_i = 0 \quad \text{in } \Omega, \tag{2.1}$$
$$\mathbf{u} \in W_0^{1,2}(\Omega; R^n).$$

In any case we suppose that the basis chosen is orthonormal in the space $L^2(\Omega; R^n)$.

DEFINITION 2.1. A weak solution to the mixed problem

$$\ddot{u}_i - \frac{\partial}{\partial x_j}\sigma_{ij}^I - \frac{\partial}{\partial x_j}\sigma_{ij}^M = f_i \quad \text{in } Q = \Omega \times (0, \infty), \tag{2.2}$$

$$\mathbf{u} = 0 \quad \text{on } \partial\Omega \times (0, \infty), \tag{2.3}$$

$$\mathbf{u}(0, \cdot\,) = \mathbf{u}^0, \qquad \ddot{\mathbf{u}}(0, \cdot\,) = \mathbf{u}^1 \tag{2.4}$$

with $u^0 \in W_0^{1,2}(\Omega; \mathbb{R}^n)$, $\mathbf{u}^1 \in L^2(\Omega; \mathbb{R}^n)$, is a function $\mathbf{u} \in L^\infty((0, \infty); W_0^{1,2}(\Omega; \mathbb{R}^n))$, for which $\dot{\mathbf{u}} \in L^\infty((0, \infty); L^2(\Omega; R^n))$, $\ddot{u} \in L^\infty((0, \infty); W^{-1,2}(\Omega; R^n))$ and such that for almost all $T > 0$, and all $\mathbf{v} \in W_0^{1,2}(\Omega; R^n)$,

$$\int_{\Omega_r} \ddot{u}_i v_i \, d\mathbf{x} + \int_{\Omega T} \sigma_{ij}^I e_{ij}(\mathbf{v}) \, d\mathbf{x} + \int_{\Omega_T} \sigma_{ij}^M e_{ij}(\mathbf{v}) \, d\mathbf{x} = \int_{\Omega_T} f_i v_i \, d\mathbf{x}. \tag{2.5}$$

In (2.2), (2.5)

$$\mathbf{f} \in L^\infty((0, \infty); L^2(\Omega; \mathbb{R}^n)) \cap L^2((0, \infty); L^2(\Omega; R^n)), \tag{2.6}$$

while $\Omega_T = (0, T) \times \Omega$. For integer k, the $W^{k,2}$ are, of course, the Sobolev spaces of L^2 functions with square integrable derivatives up to order k, while the zero subscript denotes those functions with vanishing traces on $\partial\Omega$ up to order $k-1$; via duality $W^{-k,2} = (W_0^{k,2})^*$. Spaces of fractional derivatives will be introduced below.

If we construct Galerkin approximations of the solution of (2.2)-(2.4) of the form $\mathbf{u}^n = \sum_{k=1}^n e_i(t)\mathbf{w}^i(\mathbf{x})$, we then find that for all $\mathbf{w}^k$, $k = 1, 2, \ldots, n$ and all $t > 0$,

$$\begin{aligned} &\int_\Omega \ddot{u}_i^n w_i^k \, d\mathbf{x} + \int_n \frac{\partial W}{\partial e_{ij}}(\mathbf{e}(\mathbf{u}^n))e_{ij}(\mathbf{w}^k) \, d\mathbf{x} \\ &\quad - \int_\Omega \left(\int_0^t h(t-\tau)\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(u^n(\tau)) - \mathbf{e}(u^n(t))) \, d\tau\right) e_{ij}(\mathbf{w}^k) \, d\mathbf{x} = \int_\Omega f_i w_i^k \, d\mathbf{x}, \end{aligned} \tag{2.7}$$

$$\int_{\Omega t} u_i^n(0,\cdot)w_i^k\,d\mathbf{x}=\int_{\Omega} u_i^0 w_i^k\,d\mathbf{x}, \tag{2.8}$$

$$\int_{\Omega} \dot{u}_i(0,\cdot)w_i^k\,d\mathbf{x}=\int_{\Omega} u_i^1 w_i^k\,d\mathbf{x}. \tag{2.9}$$

The problem (2.7)-(2.9) clearly possesses a unique solution on some interval $[0, t_0]$, and for this solution we may state the following.

LEMMA 2.1. *For any T, $0\leqq T<T_\infty$, T_∞ being the maximal time of existence of the solution of* (2.7)-(2.9), *the solution $\mathbf{u}^n$ satisfies, for some $c_1>0$,*

$$\begin{aligned}&\frac{1}{2}\int_{\Omega_T}\|\dot{\mathbf{u}}^n\|^2\,d\mathbf{x}+\int_{\Omega_T}W(\mathbf{e}(\mathbf{u}^n))\,d\mathbf{x}+c_1\int_{\Omega}\int_{M_{T,\delta}}\frac{\|\mathbf{e}(\mathbf{u}^n(\tau))-\mathbf{e}(\mathbf{u}^n(t))\|^2}{|t-\tau|^{1+\nu}}\,d\mathbf{x}\,dt\,d\tau\\&\leqq\int_0^T\int_{\Omega}f_iu_i^n\,d\mathbf{x}\,dt+\frac{1}{2}\int_{\Omega}\|\dot{\mathbf{u}}^n\|^2d\mathbf{x}+\int_{\Omega}W(\mathbf{e}(\mathbf{u}^n))\,d\mathbf{x},\end{aligned}\tag{2.10}$$

where

$$M_{T,\delta}=\{(t,\tau)|0<t<T, 0<\tau<T, t-\delta<\tau<t+\delta, \delta>0\}.$$

Proof. We replace $\mathbf{w}^k$ in (2.7) by $\dot{\mathbf{u}}^n$ and then integrate the equation over $[0, T]$; there results

$$\begin{aligned}&\frac{1}{2}\int_{\Omega_T}\|\dot{\mathbf{u}}\|^2\,d\mathbf{x}-\frac{1}{2}\int_{\Omega}\|\dot{\mathbf{u}}^n(0)\|\,d\mathbf{x}+\int_{\Omega_T}W(\mathbf{e}(\mathbf{u}^n))\,d\mathbf{x}\\&\quad-\int_{\Omega}W(\mathbf{e}(\mathbf{u}^n(0)))\,d\mathbf{x}\\&\quad-\int_0^T dt\int_{\Omega}\left(\int_0^t h(t-\tau)\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(\mathbf{u}^n(\tau))-\mathbf{e}(\mathbf{u}^n(t)))\,d\tau\right)\cdot\dot{e}_{ij}(\mathbf{u}^n(t))\,d\mathbf{x}\\&\quad=\int_0^t dt\int_{\Omega}f_i\dot{u}_i^n\,d\mathbf{x}\,dt.\end{aligned}\tag{2.11}$$

Denoting by $\nu=(\nu_t,\nu_\tau)$ the outer normal to ∂M_T, where

$$M_T=\{(t,\tau)|0<t<\tau, 0<t<\tau\},$$

we compute that for some $c_1>0$,

$$\begin{aligned}&-\int_{M_T}H(t-\tau)\frac{\partial}{\partial t}(e_{ij}(\mathbf{u}^n(t))-e_{ij}(\mathbf{u}^n(\tau)))\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(u^n(\tau)-\mathbf{e}(u^n(t)))\,dt\,d\tau\\&\quad=\int_{M_T}h(t-\tau)V(\mathbf{e}(\mathbf{u}^n(\tau))-\mathbf{e}(\mathbf{u}^n(t)))\,dt\,d\tau\\&\quad=\int_{\partial M_T}h(t-\tau)\frac{\partial}{\partial t}V(\mathbf{e}(\mathbf{u}^n(\tau))-\mathbf{e}(\mathbf{u}^n(t)))\nu_t\,ds\\&\quad\quad-\int_{M_T}h(t-\tau)V(\mathbf{e}(\mathbf{u}^n(\tau))-\mathbf{e}(\mathbf{u}^n(t)))\,dt\,d\tau\\&\quad\geqq c_1\lambda\int_{M_{T,\delta}}\frac{\|\mathbf{e}(\mathbf{u}^n(\tau))-\mathbf{e}(\mathbf{u}^n(t))\|^2}{|t-\tau|^{1+\nu}}\,dt\,d\tau.\end{aligned}\tag{2.12}$$

Combining (2.11) with (2.12) we find the estimate (2.10). □

In what follows we will assume that besides (2.6), $\mathbf{f}$ satisfies the condition

$$\int_\Omega \int_0^\infty \int_0^\infty \frac{\|\mathbf{f}(\tau)-\mathbf{f}(t)\|^2}{|\tau-t|^{1+\nu}}\,dt\,d\tau\,d\mathbf{x}<\infty. \tag{2.13}$$

It follows from Lemma 2.1 that the Galerkin approximants $\mathbf{u}^n$ are defined on the entire interval $[0,\infty]$. We now make the following definition.

DEFINITION 2.2. Let B be a Banach space, $\mathbf{f}(t)\in B$, $0<t<T$, and $0<\mu<1$. Then we define

$$W^{\mu,2}((0,T);B)=\left\{\mathbf{f}\in L^2((0T);B)\,\middle|\,\int_0^T \|\mathbf{f}(t)\|_B^2\,dt+\int_0^T\int_0^T \frac{\|\mathbf{f}(t)-\mathbf{f}(\tau)\|_B^2}{|t-\tau|^{1+2\mu}} \equiv \|\mathbf{f}\|^2_{W^{\mu,2}}<\infty\right\}, \tag{2.14}$$

and we have the following.

LEMMA 2.2. *For each $T>0$, $\exists C^*(T)>0$ such that the Galerkin approximants $\mathbf{u}^n$ satisfy*

$$\|\dot{\mathbf{u}}^n\|_{W^{\nu/2,2}((0,T);W^{-1,2}(\Omega,R^n))}\leqq C^*(T). \tag{2.15}$$

Proof. Let $\boldsymbol{\phi}\in W_0^{1,2}(\Omega;\mathbb{R}^n)$, and let R_n be the projection operator mapping $W_0^{1,2}(\Omega;\mathbb{R}^n)$ to the subspace spanned by $(\mathbf{w}^1,\dots,\mathbf{w}^n)$. Then

$$\begin{aligned}\|\ddot{\mathbf{u}}(\tau)-\ddot{\mathbf{u}}^n(t)\|_{W^{-1,2}}&=\sup_{\|\boldsymbol{\phi}\|_{W^{1,2}}\leqq 1}\int_\Omega (\ddot{u}_i^n(\tau)-\ddot{u}_i(t))\phi_i\,d\mathbf{x}\\ &=\sup_{\|\boldsymbol{\phi}\|_{W^{1,2}}\leqq 1}\int_\Omega (\ddot{\mathbf{u}}_i^n(\tau)-\ddot{\mathbf{u}}_i(t))(R_n\boldsymbol{\phi})_i\,d\mathbf{x}.\end{aligned} \tag{2.16}$$

We now employ (2.7) so as to estimate the right-hand side of (2.16); it is sufficient to look at the term in (2.7) generated by the memory portion of the stress σ_{ij}^M, as the estimates for the other terms follow in a straightforward fashion. We have

$$\begin{aligned}&\sup_{\|\boldsymbol{\phi}\|_{W^{1,2},\leqq 1}}\left|\int_\Omega e_{ij}(R_n\boldsymbol{\phi})\,d\mathbf{x}\int_0^{t_1} h(t_1-\tau_1)\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(\mathbf{u}^n(\tau_1))-\mathbf{e}(\mathbf{u}^n(t_1)))\,d\tau_1\right.\\ &\qquad\left.-\int_\Omega e_{ij}(R_n\boldsymbol{\phi})\,d\mathbf{x}\int_0^{t_2} h\left((t_2-\tau_2)\frac{\partial V}{\partial e_{ij}}\right)\mathbf{e}(\mathbf{u}^n(\tau_2)-\mathbf{e}(\mathbf{u}^n(t_2)))\,d\tau_2\right|\\ &\leqq c_1\left\{\int_\Omega\sum_{i,j=1}^n\left[\int_0^{t_1} h(t_1-\tau_1)\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(\mathbf{u}^n(\tau_1))-\mathbf{e}(\mathbf{u}^n(t_1)))\,d\tau_1\right.\right.\\ &\qquad\left.\left.-\int_0^{t_2} h(t_2-\tau_2)\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(\mathbf{u}^n(\tau_2))-\mathbf{e}(\mathbf{u}^n(t_2)))\,d\tau_2\right]^2 d\mathbf{x}\right\}^{1/2}\end{aligned} \tag{2.17}$$

for some $c_1>0$. For $0<t_2<t_1<T$, we may write the terms within the square bracket, on the right-hand side of the above estimate, in the form

$$\begin{aligned}&\int_0^{t_2} h(s)\left[\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(\mathbf{u}^n(t_1-s))-\mathbf{e}(\mathbf{u}^n(t_1)))-\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(\mathbf{u}^n(t_2-s))-\mathbf{e}((\mathbf{u}^n(t_2)))\right]ds\\ &+\int_{t_2}^{t} h(s)\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(\mathbf{u}^n(t_1-s))-\mathbf{e}(\mathbf{u}^n(t_1)))\,ds.\end{aligned} \tag{2.18}$$

Using Schwarz's inequality on the first integral in this last expression we find that

$$\int_\Omega d\mathbf{x}\int_0^T dt_1\int_0^{t_1}\frac{dt_2}{|t_1-t_2|^{1+\nu}}\int_0^{t_2}\Big[\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(\mathbf{u}^n(t_1-s))-\mathbf{e}(\mathbf{u}^n(t_1)))$$
$$-\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(\mathbf{u}^n(t_2-s)-\mathbf{e}(\mathbf{u}^n(t_2-s)-\mathbf{e}(\mathbf{u}^n(t_2)))\Big]^2\,ds$$

$$\leqq c_2(T)\int_\Omega d\mathbf{x}\int_0^T dt_1\int_0^{t_1}\frac{dt_2}{|t_1-t_2|^{1+\nu}}\int_0^{t_2}[\|\mathbf{e}(\mathbf{u}^n(t_1-s))-\mathbf{e}(\mathbf{u}^n(t_2-s))\|^2 \tag{2.19}$$
$$+\|\mathbf{e}(\mathbf{u}^n(t_1))-\mathbf{e}(\mathbf{u}^n(t_2))\|^2]\,ds$$
$$\leqq c_3(T)\int_0^T\int_0^T\int_\Omega\frac{\|\mathbf{e}(\mathbf{u}^n(t_1))-\mathbf{e}(\mathbf{u}^n(t_2))\|^2}{|t_1-t_2|^{1+\nu}}\,dt_1\,dt_2\,d\mathbf{x},$$

while for the second integral in (2.18) we have the estimates

$$\left[\int_{t_2}^{t_1}h(s)\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(\mathbf{u}^n(t_1-s))-\mathbf{e}(\mathbf{u}^n(t_1)))\,ds\right]^2$$
$$\leqq c_4(T)\int_{t_2}^{t_1}h^2(s)\,ds\int_{t_2}^{t_1}\|\mathbf{e}(\mathbf{u}^n(t_1-s))-\mathbf{e}(\mathbf{u}^n(t_1))\|^2\,ds \tag{2.20}$$
$$\leqq c_5(T)[(t_1-t_2)+(t_1-t_2)^{1-2\nu}]\int_{t_2}^{t_1}\|\mathbf{e}(\mathbf{u}^n(t_1-s)-\mathbf{e}(\mathbf{u}^n(t_1))\|^2\,ds.$$

However,

$$\int_\Omega d\mathbf{x}\int_0^T dt_1\int_0^{t_1}\frac{dt_2}{|t_1-t_2|^{1+\nu}}[(t_1-t_2)+(t_1-t_2)^{1-2\nu}]$$
$$\times\int_{t_2}^{t_1}\|\mathbf{e}(\mathbf{u}^n(t_1-s))-\mathbf{e}(\mathbf{u}^n(t_1))\|^2\,ds \tag{2.21}$$
$$\leqq c_6(T)\sup_{0\leqq t\leqq T}\|\mathbf{u}^n(t)\|^2_{W^{1,2}}.$$

So for some $c_\eta(T)>0$,

$$\int_\Omega d\mathbf{x}\int_0^T dt_1\int_0^{t_1}\frac{dt_2}{|t_1-t_2|^{1+\nu}}\left[\int_{t_2}^{t_1}h(s)\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(\mathbf{u}^n(t_1-s))-\mathbf{e}(\mathbf{u}^n(t_1)))\,ds\right]^2 \tag{2.22}$$
$$\leqq c_\tau(T)\sup_{0\leqq t\leqq T}\|\mathbf{u}^n(t)\|^2_{W^{1,2}}.$$

Combining (2.16), (2.17), (2.18), (2.19), and (2.22), with the obvious estimates generated by the terms in (2.7), other than the one engendered by σ_{ij}^M, and using the definition (2.14), with $\mu=\nu/2$, we are led to (2.15).

3. An interpolation lemma. Let v^i be an orthonormal basis in $L^2(\Omega)$ formed by the eigenfunctions for the problem

$$\Delta v^i+\lambda v^i=0,\quad \text{in }\Omega;\qquad v^i\in W_0^{1,2}(\Omega). \tag{3.1}$$

Let $0\leqq\eta\leqq\frac{3}{2}$, $-1\leqq\theta\leqq 1$, and consider an element $v\in L^2((0,T);\ W^{-1,2}(\Omega))$, $T>0$. Then v may be expanded in a double Fourier series

$$v=\sum_{i=0}^{\infty}\sum_{j=1}^{\infty}C_{ij}h_i(t)v^i(x), \tag{3.2}$$

where $h_0(t)=1/\sqrt{T}$, $h_i(t)=\sqrt{2/T}\cos(i\pi/T)t$. We define

$$W^{\eta,2}((0,T);\,W_0^{\theta,2}(\Omega))$$

$$=\left\{u\in L^2((0,T);\,W^{-1,2}(\Omega))\,\middle|\,\sum_{i=0}^{\infty}\sum_{j=1}^{\infty}c_{ij}^2(1+i^2)^{\eta}\lambda_j^{\theta}\equiv\|\mathbf{u}\|_{\eta,\theta}^2<\infty\right\}. \tag{3.3}$$

For $1<\eta<\frac{3}{2}$, a norm equivalent to $\|\mathbf{u}\|_{\eta,\theta}$ is given by

$$\left[\|\mathbf{u}\|^2_{L^2((0,T);W^{-1,2}(\Omega))}+\int_0^T\int_0^T\frac{\|\dot{\mathbf{u}}(t)-\dot{\mathbf{u}}(\tau)\|^2_{W_0^{\theta,2}}}{|t-\tau|^{1+2(\eta-1)}}\,d\tau\,dt\right]^{1/2}. \tag{3.4}$$

For the spaces defined by (3.3) we may state the following.

LEMMA 3.1. *Let $0\leqq\alpha\leqq 1$, $0\leqq\eta<\frac{3}{2}$. Then there exists $C>0$ such that*

$$\|\mathbf{v}\|_{W^{\alpha\eta,2}((0,T);\,W^{-\alpha,2}(\Omega))}\leqq C\|\mathbf{v}\|^{1-\alpha}_{L^2((0,T);L^2(\Omega))}\|\mathbf{v}\|^{\alpha}_{W^{\eta,2}((0,T);W^{-1,2}(\Omega))}. \tag{3.5}$$

Proof. We compute directly that

$$\begin{aligned}&\|\mathbf{v}\|_{W^{\alpha\eta,2}((0,T);\,W^{-\alpha,2}(\Omega))}\\&=\sum_{i,j}c_{ij}^2(1+i^2)^{\alpha,\eta}\lambda_j^{-\alpha}\\&\leqq\sum_{i,j}c_{ij}^{2\alpha}(1+i^2)^{\alpha\eta}\lambda_j^{-\alpha}c_{ij}^{2(1-\alpha)}\\&\leqq\left(\sum_{i,j}c_{ij}^2(1+i^2)^{\eta}\lambda_j^{-1}\right)^{\alpha}\left(\sum_{i,j}c_{ij}^2\right)^{1-\alpha}.\end{aligned}$$

□

4. Existence of a weak solution. We begin by first proving that the operator $(\partial/\partial x_j)\sigma_{ij}^M$ is Lipschitz continuous, as indicated in the remarks of § 1; this result is the essential content of the following theorem.

THEOREM 4.1. *There exists $p>0$, independent of T, such that*

$$\begin{aligned}&\|\sigma_{ij}^M(\mathbf{u}^1)-\sigma_{ij}^M(\mathbf{u}^2)\|_{L^2((0,T);L^2(\Omega;R^{n^2}))}\\&\leqq\lambda p\|\mathbf{e}(\mathbf{u}^1)-\mathbf{e}(\mathbf{u}^2)\|_{L^2((0,T);L^2(\Omega;R^{n^2}))}.\end{aligned} \tag{4.1}$$

Proof. A direct computation yields the following series of estimates:

$$\begin{aligned}&\int_0^T dt\int_\Omega(\sigma_{ij}^M(\mathbf{u}^1))-(\sigma_{ij}^M(\mathbf{u}^2))^2\,d\mathbf{x}\\&\leqq d_1\int_0^T dt\int_\Omega\left[\int_0^t(h(t-\tau))[\|\mathbf{e}(\mathbf{u}^1(\tau))-\mathbf{e}(\mathbf{u}^2(\tau))\|+\|\mathbf{e}(\mathbf{u}^1(t))-\mathbf{e}(\mathbf{u}^2(t))\|]\,d\tau\right]^2d\mathbf{x}\\&\leqq d_2\int_0^T dt\int_\Omega\|\mathbf{e}(\mathbf{u}^1(t))-\mathbf{e}(\mathbf{u}^2(t))\|^2\,d\mathbf{x}\\&\qquad+d_3\int_0^T dt\int_\Omega\left(\int_0^t h(t-\tau)\,d\tau\right)\left(\int_0^t h(t+\tau)\|\mathbf{e}(\mathbf{u}^1(\tau))-\mathbf{e}(\mathbf{u}^2(\tau))\|^2\,d\tau\right)d\mathbf{x}\\&\leqq d_4\int_0^T dt\int_\Omega\|\mathbf{e}(\mathbf{u}^1(t))-\mathbf{e}(\mathbf{u}^2(t))\|^2\,d\mathbf{x}\\&\qquad+d_5\int_0^T d\tau\int_\Omega\|\mathbf{e}(\mathbf{u}^1(\tau))-\mathbf{e}(\mathbf{u}^2(\tau))\|^2\,d\mathbf{x}\int_\tau^T h(t-\tau)\,dt\\&\leqq d_6\int_0^T dt\int_\Omega\|\mathbf{e}(\mathbf{u}^1(t))-\mathbf{e}(\mathbf{u}^2(t))\|^2\,d\mathbf{x}.\end{aligned} \tag{4.2}$$

□

THEOREM 4.2 (Global existence of a weak solution). *The initial boundary value problem* (2.2)–(2.4) *possesses, under the conditions* (1.2)–(1.6), *with* $W(0)=V(0)=(\partial W/\partial e_{ij})(0)=(\partial V/\partial e_{ij})(0)=0$, $\mathbf{u}^0 \in W_0^{1,2}(\Omega;\, R^n)$, $\mathbf{u}^1 \in L^2(\Omega; \mathbb{R}^n)$, *and* (2.6), (2.13), *a weak solution on* $(0,\infty)$ *for* $\lambda p < \hat{\alpha}$ *and* $\nu < \frac{1}{2}$. *The solution* $\mathbf{u}$ *satisfies*

$$\text{(4.3a)} \qquad \text{(i)} \qquad \mathbf{u} \in L^\infty((0,\infty);\, W_0^{1,2}(\Omega;\, R^n)),$$

$$\text{(4.3b)} \qquad \text{(ii)} \qquad \dot{\mathbf{u}} \in L^\infty((0,\infty);\, L^2(\Omega; \mathbb{R}^n)),$$

$$\text{(4.3c)} \qquad \text{(iii)} \qquad \mathbf{u} \in W^{\nu/2,2}((0,\,T);\, W_0^{1,2}(\Omega;\, R^n)) \quad \forall T>0,$$

$$\text{(4.3d)} \qquad \text{(iv)} \qquad \begin{aligned} &\mathbf{u} \in W^{\nu/2,2}((0,\,T);\, W^{-1,2}(\Omega; \mathbb{R}^n)) \\ &\quad \cap L^2((0,\,T);\, W^{-1/(1+(\nu/2))}(\Omega; \mathbb{R}^n)) \quad \forall T>0. \end{aligned}$$

Remark. The condition $\lambda p < \hat{\alpha}$ yields the result that the operator $(\partial/\partial x_j)\sigma_{ij} = (\partial/\partial x_j)\sigma_{ij}^I + (\partial/\partial x_j)\sigma_{ij}^M$ is strongly monotone.

Proof. Let $1/(1+(\nu/2)) < \alpha < 1$. Then it follows from our interpolation Lemma 3.1, with $v = \dot{u}_i^n$, that $\dot{u}_i^n$ is a bounded sequence in $W^{(1+(\nu/2))\alpha,2}((0,\,T);\, W^{-\alpha,2}(\Omega))$, which, in turn, is compactly embedded in $W^{1,2}((0,\,T);\, W^{-1,2}(\Omega))$. Therefore, we may choose a subsequence $\mathbf{u}^{(n_k)}$, such that for each $T>0$, and some $\mathbf{u}$ satisfying (4.3a–d),

$$\text{(4.4a)} \qquad \mathbf{u}^{n_k} \to \mathbf{u}, \quad \text{in } L^2((0,\,T);\, W_0^{1,2}(\Omega; \mathbb{R}^n)),$$

$$\text{(4.4b)} \qquad \dot{\mathbf{u}}^{n_k} \to \dot{\mathbf{u}}, \quad \text{in } L^2((0,\,T);\, L^2(\Omega; \mathbb{R}^n)),$$

$$\text{(4.4c)} \qquad \mathbf{u}^{n_k} \to \mathbf{u}, \quad \text{in } W^{\nu/2,2}((0,\,T);\, W^{1,2}(\Omega; \mathbb{R}^n)),$$

$$\text{(4.4d)} \qquad \ddot{\mathbf{u}}^{n_k} \to \ddot{\mathbf{u}}, \quad \text{in } W^{\nu/2,2}((0,\,T);\, W^{-1,2}(\Omega; \mathbb{R}^n)),$$

$$\text{(4.4e)} \qquad \ddot{\mathbf{u}}^{n_k} \to \ddot{\mathbf{u}}, \quad \text{in } L^2((0,\,T);\, W^{-1,2}(\Omega;\, R^n)).$$

Now, let P_n be the projection operator from $L^2((0,\,T);\, W_0^{1,2}(\Omega; \mathbb{R}^n))$ to the space spanned by the vectors $c_j(t)\mathbf{w}^j(t)$, where, for $j=1,\dots,n$, the $c_j \in L^2((0,\,T))$. From (2.7) we obtain

$$\text{(4.5)} \qquad \begin{aligned} &\int_0^T \int_\Omega \ddot{u}_i^{n_k}((\ddot{u}_i^{n_k} - P_{n_k}\mathbf{u})_i)\, d\mathbf{x}\, dt \\ &\quad + \int_0^T \int_\Omega \frac{\partial W}{\partial e_{ij}}(\mathbf{e}(\mathbf{u}^{n_k}))e_{ij}(\mathbf{u}^{n_k} - P_{n_k}\mathbf{u})\, d\mathbf{x}\, dt \\ &\quad - \int_0^T \int_\Omega \left[\int_0^t h(t-\tau)\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(\mathbf{u}^{n_k}(\tau)) - \mathbf{e}(\mathbf{u}^{n_k}(t)))\, d\tau\right] \cdot e_{ij}(\mathbf{u}^{n_k} - P_{n_k}\mathbf{u})\, d\mathbf{x}\, dt \\ &= \int_0^T \int_\Omega f_i(u_i^{n_k} - (P_{n_k}\mathbf{u})_i)\, d\mathbf{x}\, dt. \end{aligned}$$

For $\mathbf{u} \in W_0^{1,2}(\Omega; \mathbb{R}^n)$, Korn's inequality, i.e.,

$$\text{(4.6)} \qquad \int_\Omega e_{ij}(\mathbf{u})e_{ij}(\mathbf{u})\, d\mathbf{x} \geqq \kappa \|\mathbf{u}\|^2_{W^{1,2}(\Omega;\mathbb{R}^n)},$$

holds for some $\kappa > 0$ (see, e.g., [15]). Now we know that as $n_k \to \infty$

$$\text{(4.7)} \qquad P_{n_k}\mathbf{u} - \mathbf{u}, \quad \text{in } L^2((0,\,T);\, W_0^{1,2}(\Omega;\, R^n)).$$

Therefore, both

$$\text{(4.8)} \qquad \int_0^T \int_\Omega \frac{\partial W}{\partial e_{ij}}(\mathbf{e}(P_{n_k}\mathbf{u}))e_{ij}(\mathbf{u}^{n_k} - P_{n_k}\mathbf{u})\, d\mathbf{x}\, dt \to 0$$

and

$$(4.9)\qquad \int_0^T \int_\Omega \left[\int_0^t h(t-\tau)\frac{\partial V}{\partial e_{ij}}(\mathbf{e}(P_{n_k}\mathbf{u})(\tau))-\mathbf{e}(P_{n_k}\mathbf{u}(t))\,d\tau\right] e_{ij}(\mathbf{u}^{n_k}-P_{n_k}\mathbf{u})\,d\mathbf{x}\,dt \to 0$$

as $n_k \to \infty$. By virtue, therefore, of (4.8), (4.9), (4.4e), and (4.1), with $\lambda p < \alpha$, we find that as $n_k \to \infty$

$$(4.10)\qquad \mathbf{u}^{n^k} \to \mathbf{u}, \text{in } L^2((0, T);\ W_0^{1,2}(\Omega; \mathbb{R}^n)),$$

and the existence of the required global weak solution follows as a direct consequence of (2.7). □

REFERENCES

[1] S. O. LONDEN, *An existence result for a Volterra equation in a Banach space*, Trans. Amer. Math. Soc., 235 (1978), pp. 285-304.

[2] M. RENARDY, *Some remarks on the propagation and non-propagation of discontinuities in linearly viscoelastic liquids*, Rheol. Acta, 21 (1982), pp. 251-254.

[3] W. J. HRUSA AND M. RENARDY, *On wave propagation in linear viscoelasticity*, Quart. Appl. Math., 43 (1985), pp. 237-254.

[4] K. B. HANNSGEN AND R. L. WHEELER, *Behavior of the solutions of a Volterra equation as a parameter tends to infinity*, J. Integral Equations, 7 (1984), pp. 229-237.

[5] W. J. HRUSA AND M. RENARDY, *On a class of quasilinear partial integrodifferential equations with singular kernels*, J. Differential Equations, 64 (1986), pp. 195-220.

[6] ———, *A model equation for viscoelasticity with a strongly singular kernel*, SIAM J. Math. Anal, 19 (1988), pp. 257-269.

[7] M. RENARDY, *Coercive estimates and existence of solutions for a model of one-dimensional viscoelasticity with a nonintegrable memory function*, J. Integral Equations Appl., to appear.

[8] M. DOI AND S. F. EDWARDS, *Dynamics of concentrated polymer systems*, J. Chem. Soc. Faraday, 74 (1978), pp. 1789-1832; 75 (1979), pp. 38-54.

[9] P. E. ROUSE, *A theory of the linear viscoelastic properties of dilute solutions of coiling polymers*, J. Chem. Phys. 21 (1953), pp. 1271-1280.

[10] B. H. ZIMM, *Dynamics of polymer molecules in dilute solutions: viscoelasticity, flow birefringince, and dielectric loss*, J. Chem. Phys, 24 (1956), pp. 269-278.

[11] H. M. LAUN, *Description of the nonlinear shear behavior of a low density polyethylene melt by means of an experimentally determined strain dependent memory function*, Rheol. Acta, 17 (1978), pp. 1-15.

[12] J. MILOTA, J. NEČAS, AND V. ŚVERAK, *On weak solutions to a viscoelasticity model*, Comment. Math. Univ. Carolin., 31 (1990), pp. 557-565.

[13] H. ENGLER, *Weak solutions of a class of quasilinear hyperbolic integro-differential equations describing viscoelastic materials*, Arch. Rational Mech. Anal., 113 (1991), pp. 1-38.

[14] B. D. COLEMAN, *On thermodynamics, strain impulses, and viscoelasticity*, Arch. Rational Mech. Anal., 17 (1964), pp. 1-46.

[15] J. NEČAS AND I. HLAVÁČEK, *Mathematical Theory of Elastic and Elastico-Plastic Bodies*, Elsevier, New York, 1981.

[16] P. D. LAX, *Development of singularities of solutions of nonlinear hyperbolic partial differential equations*, J. Math. Phys., 5 (1964), pp. 611-613.

[17] S. KLAINERMAN AND A. MAJDA, *Formation of singularities for wave equations including the nonlinear vibrating string*, Comm. Pure Appl. Math, 33 (1980), pp. 241-263.

[18] R. DIPERNA, *Convergence of approximate solutions to conservation laws*, Arch. Rational Mech. Anal., 82 (1983), pp. 27-70.

[19] T. J. R. HUGHES, T. KATO, AND T. E. MARSDEN, *Well-posed quasi-linear hyperbolic systems with applications to nonlinear elastodynamics and general relativity*, Arch. Rational Mech. Anal., 63 (1977), pp. 273-294.

[20] S. S. ANTMAN, *The influence of elasticity on analysis: modern developments*, Bull. Amer. Math. Soc., 9 (1983), pp. 267-291.

[21] J. E. MARSDEN AND T. J. R. HUGHES, *Mathematical Foundations of Elasticity*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

SIAM J. MATH. ANAL.
Vol. 24, No. 1, pp. 46–58, January 1993

004

WEAK SOLUTION TO AN EVOLUTION PROBLEM WITH A NONLOCAL CONSTRAINT*

PETER SHI[†]

Abstract. A parabolic problem is considered with a nonlocal constraint in place of one of the standard boundary conditions. Well-posedness of the problem is proved in a weighted, fractional Sobolev space with the problem data proposed in related weighted spaces. This comes as an intrinsic requirement of the problem. The proof uses an interpolation inequality for norms of fractional Sobolev spaces and is based on an interesting choice of the test function.

Key words. nonlocal condition, weighted Sobolev space

AMS(MOS) subject classifications. primary 35K; secondary 35R

1. Introduction. In this paper, we are concerned with one-dimensional parabolic equations with a nonlocal condition—the so-called energy specification. This is a linear constraint having the form

$$\int_0^b \theta(\xi,t)d\xi = E(t), \tag{1.1}$$

where $b \in (0,1]$ is a constant and $E(\cdot)$ is a given function. Coupled with a one-dimensional parabolic equation, condition (1.1) is quite different from usual semilocal boundary conditions such as

$$-\theta_x = f(\theta)\cdot\theta \quad \text{on the boundary}, \tag{1.2}$$

where f is some nonlinear functional. The latter, though difficult to handle in its own right, can always be tied into weak formulations of the problem upon integration by parts, while condition (1.1) may not be the case using the stardard method. Our main tool used in this paper is an interpolation inequality (Lemma 3.1, §.3) combined with the basic methods described in Yurchuk [11] and in Ladyzhenskaya et al. [4, Chap. 3].

Evolution problems with linear constraint (1.1) have received attention in the last twenty years. Most of the work was directed to classical solutions, and these were carried out mostly by Cannon and his co-workers. We refer the reader to Cannon et al. [2] and references therein. We mention that a mathematical model with constraint (1.1) was recently derived in Shi and Shillor [8] in the context of thermoelasticity.

In contrast to previous work devoted to classical solution, we consider weak solutions. Let $\Omega_T = \Omega \times (0,T)$ with $T > 0$ and $\Omega = (0,1)$. We shall consider the problem

$$\theta_t - \theta_{xx} = f_1 + f_{2x}, \qquad (x,t) \in \Omega_T, \tag{1.3}$$

$$\theta_x(1,t) = 0, \qquad 0 < t < T, \tag{1.4}$$

$$\theta(x,0) = \varphi(x), \qquad 0 < x < 1, \tag{1.5}$$

*Received by the editors January 18, 1992; accepted for publication (in revised form) May 27, 1992.

[†]Department of Mathematical Sciences, Oakland University, Rochester, Michigan 48309-4401.

$$\int_0^b \theta(\xi,t)d\xi = E(t), \qquad 0 < t < T. \tag{1.6}$$

Our interest lies in the well-posedness of the problem (1.3)–(1.6) in a weighted fractional Sobolev space under compatible assumptions on data. In particular, we do not assume any classical differentibility of E. For other data φ, f_1, and f_2, we only assume their membership in certain weighted L^2-spaces. In order to keep the exposition from technical cumbersomeness, we restrict our discussion to linear equations with constant coefficients, although the method is not limited to this case by itself.

The first study of evolution problems involving energy specification goes back to Cannon [1] 1963. There the author considered the problem

$$\theta_t - \theta_{xx} = 0, \qquad x > 0, \quad t > 0, \tag{1.7}$$

$$\theta(x,0) = \varphi, \qquad x > 0, \tag{1.8}$$

$$\int_0^{x(t)} \theta(x,t)\,dx = E(t), \qquad t > 0, \tag{1.9}$$

where $x(t)$ is a given function. Introducing $f \equiv \theta(0,t)$ as the unknown, it is proved in [1] that (1.7)–(1.9) is equivalent to a Volterra integral equation of the second kind for f. The existence and uniqueness of the solution is proved with the aid of the integral equation. The author also proved that if the spacial domain is finite, then the problem reduces to an integral equation of Fredholm type whose kernel is generated by the Green's function for the heat equation. Recently, problems similar to (1.3)–(1.6) but involving nonlinearities in the equation have been treated in Lin [6] and Cannon et al. [2].

Along a different line, the problem was considered by Ionkin [3], Makarov and Kulyev [5], and Yurchuk [11]. The method in [11] is most innovative. The author established the existence of a weighted strong solution to the problem

$$\theta_t - \theta_{xx} = f, \qquad (x,t) \in \Omega_T, \tag{1.10}$$

$$\theta(1,t) = 0, \qquad 0 < t < T, \tag{1.11}$$

$$\theta(x,0) = \varphi(x), \qquad 0 < x < 1, \tag{1.12}$$

$$\int_0^1 \theta(\xi,t)d\xi = 0, \qquad 0 < t < T, \tag{1.13}$$

under certain assumptions on data. Unlike most cases where a weighted space appears either because of singular coefficients of the equation or because of an unbounded domain, here the weight comes to place for the annihilation of inconvenient terms during integration by parts. The result and the method in the present paper are further elaboration of those in [11].

We now introduce appropriate function spaces. Let $L^2(\Omega_T; x)$ be the weighted L^2-space with finite norm

$$\|f\|_{L^2_x} = \left(\int_{\Omega_T} x^2 f^2\,dx\,dt\right)^{1/2}. \tag{1.14}$$

Let $W^{1,0}(\Omega_T; x)$ be the subspace of $L^2(\Omega_T)$ with finite norm

$$\|u\|_{W^{1,0}_x} \equiv \left(\int_{\Omega_T} u^2\,dx\,dt\right)^{1/2} + \left(\int_{\Omega_T} x^2 u_x^2\,dx\,dt\right)^{1/2}. \tag{1.15}$$

$W^{2,1}(\Omega_T;x)$ is the subspace of $W^{1,0}(\Omega_T;x)$ whose elements satisfy u_t, $u_{xx} \in L^2(\Omega_T;x)$. In general, a function in the space $W^{k,h}(\Omega_T;x)$, with k, h nonnegative integers, possesses x-derivatives up to kth order in $L^2(\Omega_T;x)$, and t-devivatives up to hth order in $L^2(\Omega_T;x)$. In this paper, we also use weighted spaces on the interval $\Omega = (0,1)$ such as $L^2(\Omega;x)$ and $H^1(\Omega;x)$, whose definitions are analogous to the spaces on Ω_T. For example, $H^1(\Omega;x)$ is the subspace of $L^2(0,1)$ with the finite norm

$$\|\varphi\|_{H^1_x} \equiv \left(\int_0^1 \varphi^2 dx\right)^{1/2} + \left(\int_0^1 x^2\varphi_x^2 dx\right)^{1/2}. \tag{1.16}$$

The following theorem is proved in [11].

THEOREM 1.1. *Assume that* $\varphi \in H^1(\Omega;x)$ *and* $f \in L^2(\Omega_T;x)$. *Then problem* (1.10)–(1.13) *admits a solution in* $W^{2,1}(\Omega_T;x)$.

To state the result of the present paper, we require the definition of the fractional Sobolev space $W^{1,\alpha}(\Omega_T;x)$, $0<\alpha<1$, which is a subspace of $W^{1,0}(\Omega_T;x)$ having the finite integral

$$I_w(u) = \int_{-\infty}^{+\infty}\int_0^1 x^2|w^{\sim}|^2(x,s)(1+s^2)^{\alpha}\,dx\,ds$$

for each $w \in C_0^{\infty}(R^2)$ whose restriction to Ω_T equals u. Here and in the rest of the paper, the tilde sign $(\cdot)^{\sim}$ denotes the Fourier transform of the underlying function. The norm of $W^{1,\alpha}(\Omega_T;x)$ is given by

$$\|u\|_{W^{1,\alpha}(\Omega_T;x)} = \|u\|_{W^{1,0}(\Omega_T;x)} + \inf_w I_w(u). \tag{1.17}$$

The space $W^{0,\alpha}(\Omega_T;x)$ and $H^{\alpha}(0,T)$ are analagously defined. We refer the reader to [7] and [9] for more details on fractional Sobolev spaces.

We are now in a position to state the main result of this paper.

THEOREM 1.2. *Assume that* f_1, $f_2 \in L^2(\Omega_T;x)$ *with* f_2 *continuous in a neighborhood of* $x=1$, $\varphi \in L^2(\Omega;x)$, *and* $E \in H^{(1+\epsilon)/2}(0,T)$, *where* $\epsilon > 0$. *Then problem* (1.3)–(1.6) *admits a unique solution* θ *in* $W^{1,1/2}(\Omega_T;x)$, *provided that, in addition, one of the following conditions holds*:

(i) $b=1$;

(ii) $\varphi \in H^1(b-\sigma,1)$ *and* $f_{2x} \in L^2(b-\sigma,1)$ *for some small* $\sigma > 0$.

Moreover, we have the estimates

$$C\,\|\theta\|_{W_x^{1,1/2}} \le \begin{cases} \Sigma & \textit{if } b=1,\\ \Sigma + \|f_{2x}\|_{L^2(\Omega_{\sigma,T})} + \|\varphi\|_{H^1(b-\sigma,1)} & \textit{if } 0<b<1,\end{cases}$$

where $\Sigma = \|f_1\|_{L^2_x} + \|f_2\|_{L^2_x} + \|\varphi\|_{L^2_x} + \|E\|_{H^{\alpha}(0,T)}$, $\Omega_{\sigma,T} = (b-\sigma,1)\times(0,T)$, *and* $C>0$ *depends* ϵ, α, σ, b, *and* T *only.*

2. Weak formulation. In general, a $W^{1,1/2}(\Omega_T;x)$ solution to problem (1.3)–(1.6) can be understood in sense of distributions and traces, but we prefer a variational formulation in which the nonlocal constraint (1.6) does not explicitly appear. Following the idea of [11], we introduce the operators

$$J(v) \equiv \int_x^1 v(\xi,t)\,d\xi, \tag{2.1}$$

$$M(v) = x^2 v(x,t) + 2xJ(v). \tag{2.2}$$

For any function $v \in C^1(\overline{\Omega}_T)$, it is easy to check that the following holds:

$$J(v)(1,t) = 0, \qquad M(v)(0,t) = 0, \tag{2.3}$$

and

$$2\int_0^1 xvJ(v)\,dx = \int_0^1 |J(v)|^2\,dx \tag{2.4}$$

for all $t \in [0,T]$. These equalities can be extended to functions in $W^{1,0}(\Omega_T; x)$ by a density argument. Let $Lu = u_t - u_{xx}$. We next compute the integral $\int_{\Omega_T}(Lu)\cdot M(v)\,dx$. We first work with the special case, where

$$b = 1 \quad \text{and} \quad E(t) = 0 \quad \forall t \in (0,T). \tag{2.5}$$

Thus in the following computations we assume $u, v \in C^1(\overline{\Omega}_T)$, $u_x(1,t) = 0$, $u(x,0) = \varphi(x)$, and both u and v satisfy the homogeneous integral condition

$$\int_0^1 u(\xi,t)\,d\xi = \int_0^1 v(\xi,t)\,d\xi = 0. \tag{2.6}$$

In light of the above assumptions, we obtain

$$\begin{aligned} -\int_0^1 x^2 u_{xx} v\,dx &= -\left\{ x^2 v u_x \Big|_{x=0}^{x=1} - \int_0^1 u_x[2xv + x^2 v_x]\,dx \right\} \\ &= \int_0^1 2xvu_x\,dx + \int_0^1 x^2 u_x v_x\,dx, \end{aligned} \tag{2.7}$$

$$\begin{aligned} -2\int_0^1 u_{xx} xJ(v)\,dx &= -2\left\{ xJ(v)u_x \Big|_{x=0}^{x=1} - \int_0^1 u_x[J(v) - xv]\,dx \right\} \\ &= 2\int_0^1 u_x J(v)\,dx - 2\int_0^1 xvu_x\,dx \\ &= 2\left\{ J(v)u\Big|_{x=0}^{x=1} + \int_0^1 vu\,dx \right\} - 2\int_0^1 xvu_x\,dx \\ &= 2\int_0^1 uv\,dx - 2\int_0^1 xvu_x\,dx. \end{aligned} \tag{2.8}$$

It follows from (2.7) and (2.8) that

$$\int_0^1 -u_{xx}M(v)\,dx = \int_0^1 x^2 u_x v_x\,dx + 2\int_0^1 uv\,dx. \tag{2.9}$$

To compute the integral $u_t M(v)$, we assume in addition, $v(\cdot,T) = 0$. The standard integration by parts leads to

$$\int_0^T u_t M(v)\,dt = -\varphi(x)M\big(v(x,0)\big) - \int_0^T uM(v_t)\,dt. \tag{2.10}$$

Thus from (2.9) and (2.10) we obtain

$$\begin{aligned} &\int_{\Omega_T} (Lu)M(v)\,dx\,dt + \int_0^1 \varphi M\big(v(x,0)\big)\,dx \\ &= \int_{\Omega_T} [-uM(v_t) + x^2 u_x v_x + 2uv]\,dx\,dt. \end{aligned} \tag{2.11}$$

Returning to the general case, without assuming (2.5), we introduce the transform

$$u = \theta - E - \int_b^1 \theta\, d\xi, \tag{2.12}$$

where θ is a classical solution of (1.3)–(1.6). It is easy to check that u thus defined in (2.12) satisfies (2.6) and other conditions that lead to (2.11). Moreover, u satisfies the equation

$$Lu = f_1 + f_{2x} - \left(E + \textstyle\int_b^1 \theta\, d\xi\right)_t . \tag{2.13}$$

Substituting (2.12) and (2.13) into (2.11), after appropriate integration by parts, yields

$$\begin{aligned}\int_{\Omega_T} &\left[-\left(\theta - E - \textstyle\int_b^1 \theta\, d\xi\right) M(v_t) + x^2\theta_x v_x + 2\left(\theta - E - \textstyle\int_b^1 \theta\, d\xi\right) v\right] dx\, dt \\ &= \int_{\Omega_T} \left[f_1 M(v) + f_2\left(x^2 v_x + 2J(v)\right) + \left(E + \textstyle\int_b^1 \theta\, d\xi\right) M(v_t)\right] dx\, dt \\ &\quad + \int_0^1 \left[\varphi - E(0) - \textstyle\int_b^1 \varphi(\xi)\, d\xi\right] M(v(x,0))\, dx + \int_0^T f_2(1,t) v(1,t)\, dt,\end{aligned}$$

which in turn, by virtue of a straightforward simplification, gives

$$\begin{aligned}\int_{\Omega_T} &\left[-\theta M(v_t) + x^2\theta_x v_x + 2\left(\theta - E - \textstyle\int_b^1 \theta\, d\xi\right) v\right] dx\, dt \\ &= \int_{\Omega_T} \left[f_1 M(v) + f_2\left(x^2 v_x + 2J(v)\right)\right] dx\, dt \int_0^T f_2(1,t) v(1,t)\, dt \\ &\quad + \int_0^1 \left[\varphi - E(0) - \textstyle\int_b^1 \varphi(\xi)\, d\xi\right] M(v(x,0))\, dx.\end{aligned} \tag{2.14}$$

Definition 2.1. We say that $\theta \in W^{1,0}(\Omega_T)$ is a weak solution to problem (1.3)–(1.6) if (2.14) holds for all $v \in C^1(\overline{\Omega_T})$ such that

$$v(\cdot, T) = 0 \quad \text{and} \quad \int_0^1 v(\xi, \cdot)\, d\xi = 0. \tag{2.15}$$

The above computations leading to (2.14) have shown that any classical solution to (1.3)–(1.6) is also a weak solution. Reversing the steps above, we can also prove that a sufficiently smooth weak solution is a classical solution.

We remark that each term in (2.14) is well defined if the problem data satisfies the requirement of Theorem 1.2.

3. A priori estimates. In this section we derive estimates for classical solutions to problem (1.3)–(1.6) in the $W^{1,1/2}(\Omega_T; x)$ norm. The existence of a weak solution is obtained by subtracting a convergent subsequence, once a priori estimates are established. Regularity assumptions on problem data are not important for our purpose in this section since they can be approximated by smooth functions.

We use θ to denote the smooth solution to (1.3)–(1.6), and u is determined by the transform (2.12). These conventions will be in force throughout this section unless otherwise specified.

We first prove an interpolation inequality, which plays a crucial role in the subsequent estimates. The inequality may also be of interest by itself.

LEMMA 3.1. *For any $\frac{1}{2} < \alpha < 1$, there exists a constant $C_\alpha > 0$ such that for any $w, v \in C^1[0,T]$,*

$$\left|\int_0^T w'v\,dt\right| \le C_\alpha \|w\|_{H^\alpha(0,T)} \cdot \|v\|_{H^{1-\alpha}(0,T)}. \tag{3.1}$$

For any $0 < \beta < 1$ and $\delta > 0$, there exists $C_{\delta,\beta} > 0$ such that

$$\|v\|_{H^\beta(0,T)} \le \delta\|v\|_{H^1(0,T)} + C_{\delta,\beta}\|v\|_{L^2(0,T)} \quad \forall\ v \in H^1(0,T). \tag{3.1'}$$

Proof. A simple proof of (3.1′) can be made by contrapositive argument, using the fact of compact imbeddings

$$H^1(0,T) \hookrightarrow H^\beta(0,T) \hookrightarrow L^2(0,T).$$

Since (3.1′) is essentially known, we omit the proof.

To prove (3.1) we choose $f \in H^\alpha(R), g \in H^{1-\alpha}(R)$ such that $f|_{(0,T)} = w, g|_{(0,T)} = v$. Fix $\tau < T$. Then for all $h > 0$ sufficiently small, we get by the Plancherel theorem that

$$\begin{aligned} I_h &= \int_0^\tau \frac{w(t+h) - w(t)}{h} v(t)\,dt \\ &= \int_0^\tau \frac{f(t+h) - f(t)}{h} g(t)\,dt \\ &= \int_{-\infty}^\infty \frac{f(t+h) - f(t)}{h} (\chi_\tau g)(t)\,dt \\ &= \int_{-\infty}^\infty \frac{e^{ihs} - 1}{h} \tilde{f}(s)(\chi_\tau g)^\sim(s)\,ds, \end{aligned} \tag{3.2}$$

where χ_τ is the characteristic function of $(0,\tau)$. Now

$$\left|\frac{e^{ihs} - 1}{h}\right| \le 2|s| \quad \forall h > 0,$$

and

$$\lim_{h\to 0} \frac{e^{ihs} - 1}{h} = is \quad \forall s \in R.$$

Thus for all $h > 0$ the modulus of the integrand in the last integral in (3.2) is dominated by

$$k(s) = 2|s|\ |\tilde{f}(s)|\ |(\chi_\tau g)^\sim(s)|,$$

and $k \in L^1(R)$ since $f \in H^\alpha(R)$, $g \in H^{1-\alpha}(R)$. Hence by the dominated convergence theorem,

$$\begin{aligned} \left|\int_0^\tau u'v\,dt\right| &= \lim_{h\to 0} |I_h| \\ &= \left|\int_{-\infty}^\infty is\tilde{f}(s)(\chi_\tau g)^\sim(s)\,ds\right| \\ &\le \int_{-\infty}^\infty (1 + |s|^2)^{\alpha/2}|\tilde{f}(s)|\,(1 + |s|^2)^{(1-\alpha)/2}|(\chi_\tau g)^\sim(s)|\,ds \\ &\le \|f\|_{H^\alpha(R)} \cdot \|\chi_\tau g\|_{H^{1-\alpha}(R)} \\ &\le C\|f\|_{H^\alpha(R)} \cdot \|g\|_{H^{1-\alpha}(R)}, \end{aligned}$$

where the last inequality follows from the fact that multiplication by χ_τ is a bounded operator on $H^{1-\alpha}(R)$ with operator norm bounded by a constant $C > 0$ that does not depend on τ (Strichartz [9, Cor. II. 3.7]). Thus by definition of the norm in $H^\alpha(0,T)$ we get that

$$\left| \int_0^\tau w'v \, dt \right| \leq C \|w\|_{H^\alpha(0,T)} \|v\|_{H^{1-\alpha}(0,T)}.$$

The result follows by letting $\tau \to T$.

LEMMA 3.2. *For each $\frac{1}{2} < \alpha < 1, \epsilon > 0$, there exists a constant $C > 0$ that depends only on α and ϵ such that*

$$\|u\|_{W_x^{1,0}} \leq C \left\{ \|f_1\|_{L_x^2} + \|f_2\|_{L_x^2} + \|E\|_{H^\alpha(0,T)} + \|\varphi\|_{L_x^2} + \|\textstyle\int_b^1 \theta \, d\xi\|_{H^\alpha(0,T)} \right\}$$
$$+ \epsilon \|u\|_{W_x^{0,1-\alpha}}.$$

Proof. By virtue of (2.9) in which we set $v = u$, we obtain

$$\int_0^1 -u_{xx} M(u) \, dx = \int_0^1 x^2 u_x^2 \, dx + 2 \int_0^1 u^2 \, dx.$$

Adding to both sides of the above equality by $\int_0^1 u_t M(u) dx$ and integrating the result over $(0,T)$, give

$$\int_{\Omega_T} \left(u_t \, M(u) + x^2 u_x^2 + 2u^2 \right) dx \, dt = \int_{\Omega_T} Lu \cdot M(u) \; dx \, dt.$$

Invoking (2.12) and (2.13), and carrying out appropriate integration by parts, we obtain

$$\begin{aligned} &\int_{\Omega_T} \left(u_t \, M(u) + x^2 u_x^2 + 2u^2 \right) dx \, dt \\ &\quad = \int_{\Omega_T} \left[f_1 \, M(u) + f_2 \big(x^2 u_x + 2J(u) \big) + F_t M(u) \right] dx \, dt, \end{aligned} \tag{3.3}$$

where

$$F \equiv -E - \textstyle\int_b^1 \theta \, d\xi. \tag{3.4}$$

The first term of the integral on the left-hand side of (3.3) is controlled from below by the integral

$$\int_{\Omega_T} \left[2xu_t \int_x^1 u(\xi, t) \, d\xi \right] dx \, dt - \frac{1}{2} \int_0^1 x^2 \varphi^2 \; dx.$$

Therefore, using the definition for the $W^{1,0}(\Omega_T; x)$ norm, we infer from (3.3),

$$\begin{aligned} &\int_{\Omega_T} \left[2xu_t \int_x^1 u(\xi, t) \, d\xi \right] dx \, dt + C_1 \|u\|_{W_x^{1,0}}^2 \\ &\quad \leq C_2 \{ \|f_1\|_{L_x^2}^2 + \|f_2\|_{L_x^2}^2 + \|\varphi\|_{L_x^2}^2 \} + \left| \textstyle\int_{\Omega_T} F_t \, M(u) \, dx \, dt \right|, \end{aligned} \tag{3.5}$$

where $C_1 > 0$, $C_2 > 0$, are fixed constants. The calculations from (3.3) to (3.5) are straightforward but somewhat tedious. We only illustrate a typical step below.

$$\begin{aligned} 2\int_{\Omega_T} f_2 J(u)\,dx\,dt &= -2\int_0^T\int_0^1 J(u)\,dJ(f_2)\,dx\,dt \\ &= -2\int_0^T\int_0^1 uJ(f_2)\,dx\,dt \\ &\le \frac{1}{4\epsilon}\int_0^T\int_0^1 J^2(f_2)\,dx\,dt + \epsilon\int_{\Omega_T} u^2\,dx\,dt, \end{aligned}$$

where $\epsilon > 0$ is sufficient small. In light of the inequality

$$\int_0^1 J^2(f_2)\,dx \le 4\int_0^1 x^2 f_2^2\,dx,$$

the term $2\int_{\Omega_T} f_2 J(u)\,dx\,dt$ in (3.4) is then dominated by

$$\frac{1}{\epsilon}\int_{\Omega_T} x^2 f_2^2\,dx\,dt + \epsilon\int_{\Omega_T} u^2\,dx\,dt, \tag{3.6}$$

where the second term will be absorbed in the left-hand side of (3.3).

We now estimate the integrals on both sides of (3.5). In light of Lemma 3.1, we have

$$\left|\int_{\Omega_T} F_t M(u)\,dx\,dt\right| \le \frac{1}{4\epsilon}\|F\|^2_{H^\alpha(0,T)} + \epsilon\|u\|^2_{W_x^{0,1-\alpha}}, \tag{3.7}$$

where $\epsilon > 0$ is sufficiently small. The integral on the left-hand side of (3.5) can be estimated in virtue of the following identity:

$$\int_0^1 x\big[uJ(v) + vJ(u)\big]\,dx = \int_0^1 J(u)J(v)\,dx \quad \forall u, v \in C[0,1]. \tag{3.8}$$

For this, we assume, without loss of generality, $f_2(1,\cdot) = 0$. Indeed, replacing u by u_t, v by u, in (3.8), and integrating the result over $[0,T]$, yield

$$\begin{aligned} &\int_{\Omega_T}\left[2xu_t\int_x^1 u(\xi,t)\,d\xi\right]dx \\ &\quad= 2\int_{\Omega_T} J(u_t)J(u)\,dx\,dt - 2\int_{\Omega_T} xu\,J(u_t)\,dx\,dt \\ &\quad\ge -\int_0^1 J^2(\varphi)\,dx - 2\int_0^T\int_0^1 xu\left[\int_x^1 (u_{\xi\xi} + f_1 + f_{2\xi} + F_t)\,d\xi\right]dx\,dt \\ &\quad= -\int_0^1 J^2(\varphi)\,dx - 2\int_0^T\int_0^1 xu\left[-u_x + \int_x^1 f_1\,d\xi - f_2(x) + \int_x^1 F_t\,d\xi\right]dx\,dt \\ &\quad\ge -\int_0^1 J^2(\varphi)\,dx - C_3\left\{\|f_1\|^2_{L_x^2} + \|f_2\|^2_{L_x^2} + \frac{1}{4\epsilon}\|F\|^2_{H^\alpha(0,T)} + \epsilon\|u\|_{W_x^{0,1-\alpha}}\right\}. \end{aligned} \tag{3.9}$$

Substituting (3.7) and (3.9) into (3.3)–(3.5), we obtain the desired result.

The next lemma improves the estimates in Lemma 3.2.

LEMMA 3.3. *For each $\frac{1}{2} < \alpha < 1$ there exists a constant $C = C(\alpha) > 0$, such that*

$$(3.10)\quad \|u\|_{W_x^{1,1/2}} \le C\Big\{\|f_1\|_{L_x^2} + \|f_2\|_{L_x^2} + \|\varphi\|_{L_x^2} + \|E\|_{H^\alpha(0,T)} + \|\ \textstyle\int_b^1 \theta d\xi\ \|_{H^\alpha(0,T)}\Big\}.$$

Proof. Following the method in Ladyzhenskaya et al. [4, p. 161] we extend the domain of u first to $(0,1) \times (0,+\infty)$ such that u vanishes for sufficiently large t. Then we extend the function to $Q \equiv (0,1) \times (-\infty,+\infty)$ by reflection. More precisely, once u is extended to $(0,1) \times (0,\infty)$, we set

$$u_1 = \begin{cases} u(x,t), & t \ge 0, \\ u(x,-t), & t < 0. \end{cases}$$

For the derivative of u we introduce a similar extension, namely,

$$u_2 = \begin{cases} u_x(x,t), & t \ge 0, \\ -u_x(x,-t), & t < 0. \end{cases}$$

The minus sign is designated to coping with the differential equations. In the same way, the problem data are extended as follows.

$$f_1^* = \begin{cases} f_1(x,t), & t \ge 0, \\ -f_1(x,-t), & t < 0. \end{cases}$$
$$f_2^* = \begin{cases} f_{2x}(x,t), & t \ge 0, \\ -f_{2x}(x,-t), & t < 0. \end{cases}$$
$$F^* = \begin{cases} F(x,t), & t \ge 0, \\ F(x,-t), & t < 0. \end{cases}$$

With the above extensions, (2.13) reads

$$(3.11)\qquad u_{1t} - u_{2x} = f_1^* + f_2^* + F_t^* \quad \text{a.e. in } Q = (0,1) \times (-\infty + \infty).$$

Let $\delta > 0$ and choose any $w \in C^\infty(-\infty,+\infty)$ such that $w = 1$ on $[0,T]$ and $\operatorname{supp}(w) \subset [-\delta,\, T+\delta]$. Then (3.11) implies

$$\begin{aligned} (w\,u_1)_t &= w_t\,u_1 + w\,u_{1t} \\ (3.12)\qquad &= w_t u_1 + (wu_2)_x + wf_1^* + wf_2^* + (wF^*)_t - w_tF^*. \end{aligned}$$

We now let ψ be any function in $C^1[0,1]$, possibly complex valued, such that

$$(3.13)\qquad \psi_x(1) = 0, \qquad \int_0^1 \psi(x)\,dx = 0.$$

Multiplying on both sides of (3.12) by $M(\psi)$ and integrating each term over $[0,1]$ with respect to x, taking care of $\int_0^1 (wu_2)_x M(\psi) dx$ via integration by parts, yield

$$\begin{aligned} \int_0^1 (wu_1)_t\, M(\psi)\, dx &= \int_0^1 (w_tu_1)\, M(\psi)\, dx + \int_0^1 wu_2\big(x^2\psi_x + 2J(\psi)\big)\, dx \\ (3.14)\qquad &\quad + \int_0^1 f_1^* w\, M(\psi)\, dx + \int_0^1 f_2^* w\, M(\psi)\, dx \\ &\quad + \int_0^1 (wF^*)_t\, M(\psi)\, dx - \int_0^1 w_tF^*\, M(\psi). \end{aligned}$$

Applying the Fourier transform on both sides of (3.14), we obtain

$$\int_0^1 -is(wu_1)^\sim(x,s)\,M(\psi)\,dx = Z^\sim(s) + \int_0^1 -is(wF^*)^\sim(x,s)\,M(\psi)\,dx \tag{3.15}$$

where

$$\begin{aligned} Z^\sim(s) = &\int_0^1 \left(w_t u_1 + f_1^* w + f_2^* - w_t F^*\right)^\sim M(\psi)\,dx \\ &+ \int_0^1 (wu_2)^\sim(x,s)\left(x^2\psi_x + 2J(\psi)\right)dx. \end{aligned} \tag{3.16}$$

In (3.14), (3.15) we put

$$\psi(x) = \begin{cases} i \text{ complex conjugate of } (wu_1)^\sim(x,s), & \text{if } s \geq 0, \\ -i \text{ complex conjugate of } (wu_1)^\sim(x,s), & \text{if } s < 0. \end{cases} \tag{3.17}$$

Clearly such choice of ψ satisfies (3.13). We integrate the result with respect to s to obtain

$$\begin{aligned} &\int_Q |s|\,(wu_1)^\sim M\left[\overline{(wu_1)^\sim}\right] dx\,ds \\ &\quad = \int_{-\infty}^{+\infty} Z^\sim(s)\,ds + \int_Q |s|\,(wF^*)^\sim(x,s)\,M\left[\overline{(wu_1)^\sim}\right] dx\,ds. \end{aligned} \tag{3.18}$$

The integral on the left-hand side of (3.18) is equal to, by virtue of the identity (2.4) and the Plancherel theorem [10],

$$\int_Q |s|\,x^2|(wu_1)^\sim|^2\,dx\,ds + \int_Q |s|\;|J\left((wu_1)^\sim\right)|^2\,dx\,ds. \tag{3.19}$$

Using the Plancherel theorem again and integration by parts, as in the proof of Lemma 3.2, the first integral on the right-hand side of (3.18) can be estimated by a constant times

$$\int_Q |wu_1|^2\,dx\,dt + \int_Q |wu_2|^2\,dx\,dt + \|F\|_{L^2(0,T)} + N(f_1,f_2,\varphi), \tag{3.20}$$

where $N(f_1,f_2,\varphi)$ is bounded as long as $\{f_1,f_2,\varphi\}$ remain bounded in their $L^2(\Omega_T;x)$ and $L^2(\Omega;x)$ norms, respectively. It then follows from (3.18)–(3.20) that

$$\begin{aligned} &\int_Q |s|\,x^2|(wu_1)^\sim|^2\,dx\,dt \\ &\quad \leq C\int_Q \left(|wu_1|^2 + x^2|wu_2|^2\right)dx\,dt + \|F\|_{L^2(0,T)} + N(f_1,f_2,\varphi) \\ &\qquad + \left|\int_Q |s|\,(wF^*)^\sim(x,s)\,M\left[\overline{(wu_1)^\sim}\right] dx\,ds\right|, \end{aligned} \tag{3.21}$$

where $C > 0$ depends only on the choice of w. In light of the definitions for u_1 and u_2, the first integral on the right-hand side of (3.21) is controlled by $C\|u\|^2_{W^{1,0}(\Omega_T;x)}$

for some constant $C > 0$, depending only on w. The left-hand side of (3.21) is greater or equal to $\|u\|^2_{W^{1,1/2}(\Omega_T;x)}$ by the definition of the fractional Sobolev space. Based on these estimates, (3.21) implies

$$\begin{aligned} \|u\|^2_{W_x^{1,1/2}} \leq & C\|u\|^2_{W^{1,0}(\Omega_T;x)} + N(f_1, f_2, \varphi) \\ & + \left| \int_0^1 |s|\, x^2 |(wF^*)^\sim(x,s)\, M\left[\overline{(wu_1)^\sim}\right] dx\, ds \right| \end{aligned} \tag{3.22}$$

Finally, we estimate the right-hand side of (3.22) as follows:

$$\begin{aligned} & \left| \int_Q |s|\, (wF^*)^\sim(x,s)\, M\left[\overline{(wu_1)^\sim}\right] dx\, ds \right| \\ & \leq \left| \int_Q \left[(1+|s|^2)^{\alpha/2}(wF^*)^\sim\right](1+|s|^2)^{(1-\alpha)/2} M\left[\ \overline{(wu_1)^\sim}\ \right] dx\, ds \right| \\ & \leq \frac{1}{4\epsilon}\int_Q (1+s^2)^\alpha |(wF^*)^\sim|^2\, dx\, ds + \epsilon \int_Q (1+s^2)^{1-\alpha} \left|(wu_1)^\sim\right|^2 dx\, ds \\ & \leq \frac{1}{2\epsilon}\|F\|^2_{H^\alpha(0,T)} + 2\epsilon\|u\|^2_{W_x^{0,1-\alpha}} + N(f_1, f_2, \varphi), \end{aligned} \tag{3.23}$$

where in the last inequality it is necessary to choose w properly so that

$$\int_Q (1+s^2)^\alpha |(wF^*)^\sim|^2\, dx\, ds \leq 2\|F\|^2_{H^\alpha(0,T)} + N(f_1, f_2, \varphi).$$

We now conclude from (3.22), (3.23), and Lemma 3.2 that

$$\|u\|^2_{W_x^{1,1/2}} \leq N(f_1, f_2, \varphi) + \|F\|_{H^\alpha(0,T)} + \epsilon\|u\|^2_{W_x^{0,1-\alpha}}. \tag{3.24}$$

In view of (3.3), Lemma 3.3 follows by choosing $0 < \epsilon < 1$ in (3.24).

Proof of Theorem 1.2. The result of Theorem 1.2 follows almost immediately from Lemma 3.3. If $b = 1$, then (3.10) and (2.12) imply

$$\|\theta\|_{W_x^{1,1/2}} \leq C\left\{ \|f_1\|_{L_x^2} + \|f_2\|_{L_x^2} + \|\varphi\|_{L_x^2} + \|E\|_{H^\alpha(0,T)} \right\}. \tag{3.25}$$

Consequently, a weak solution in $W^{1,1/2}(\Omega_T; x)$ is obtained by extracting a convergent subsequence of these classical solutions and taking the limit in (2.14). Since the problem is linear, uniqueness of the solution follows from (3.25). The case when $0 < b < 1$ needs a bit more effort. Instead of (3.25), we have

$$\begin{aligned} \|\theta\|_{W_x^{1,1/2}} \leq C\Big\{ & \|f_1\|_{L_x^2} + \|f_2\|_{L_x^2} + \|\varphi\|_{L_x^2} + \|E\|_{H^\alpha(0,T)} \\ & + \|f_{2x}\|_{L^2(\Omega_{\sigma,T})} + \|\varphi\|_{H^1(b-\sigma,1)}. \Big\}, \end{aligned} \tag{3.26}$$

where $\Omega_{\sigma,T} = (b-\sigma, 1) \times (0,T)$. To derive (3.26) we employ cutoff functions. Let $w \in C^2[0,1]$ such that $w = 1$ on $[b,1]$ and $w = 0$ on $[0, b-\sigma]$. Let

$$\Theta = w(x)\theta(x,t)e^{-\lambda t}, \qquad \lambda > 0.$$

By straightforward computations using (1.3)–(1.6), we obtain

$$\Theta_t - \Theta_{xx} + \lambda\Theta = fe^{-\lambda t}, \qquad (x,t) \in \Omega_{\sigma,T}, \tag{3.27}$$

$$\Theta_x(1,t) = 0, \qquad 0 < t < T, \tag{3.28}$$

$$\Theta(b-\sigma,t) = 0, \qquad 0 < t < T, \tag{3.29}$$

$$\Theta(x,0) = w\varphi, \qquad b-\sigma < x < 1, \tag{3.30}$$

where

$$f = \big(f_1 + f_{2x}\big)w - \big(2\theta_x w_x + \theta w_{xx}\big).$$

Following standard $W^{2,1}(\Omega_{\sigma,T})$ estimates, squaring both sides of (3.27) and integrating by parts, we obtain

$$\|\Theta\|_{W^{2,1}(\Omega_{\sigma,T})} + \lambda^2\|\Theta\|_{L^2(\Omega_{\sigma,T})} \le C\|f\|_{L^2(\Omega_{\sigma,T})}, \tag{3.31}$$

where $C > 0$ is independent of f and λ. In light of Lemma 3.3, f can be further controlled by

$$\begin{aligned} \|f\|_{L^2(\Omega_{\sigma,T})} \le& \|f_1\|_{L^2(\Omega_{\sigma,T})} + \|f_{2x}\|_{L^2(\Omega_{\sigma,T})} \\ &+ \|\varphi\|_{H^1(b-\sigma,1)}\|E\|_{H^\alpha(0,T)} + C\left\|\int_b^1 \theta\, d\xi\right\|_{H^\alpha(0,T)}. \end{aligned} \tag{3.32}$$

Now applying (3.1′) of Lemma 3.1 to $\int_b^1 \theta(\xi,\cdot)\, d\xi$ yields

$$\begin{aligned} \left\|\int_b^1 \theta\, d\xi\right\|_{H^\alpha(0,T)} &\le \delta\left\|\int_0^1 \theta_t\, d\xi\right\|_{L^2(0,T)} + C_\delta\left\|\int_b^1 \theta\, d\xi\right\|_{L^2(0,T)} \\ &\le \delta\|\Theta\|_{W^{2,1}(\Omega_{\sigma,T})} + C_\delta\|\Theta\|_{L^2(\Omega_{\sigma,T})}, \end{aligned} \tag{3.33}$$

where $\delta > 0$ is arbitrary and $C_\delta > 0$. It follows from (3.31)–(3.33) that, by setting δ sufficiently small and λ sufficiently large,

$$\begin{aligned} &\|\Theta\|_{W^{2,1}(\Omega_{\sigma,T})} \\ &\le C\Big\{\|f_1\|_{L^2(\Omega_{\sigma,T})} + \|f_{2x}\|_{L^2(\Omega_{\sigma,T})} + \|\varphi\|_{H^1(b-\sigma,1)} + \|E\|_{H^\alpha(0,T)}\Big\}, \end{aligned} \tag{3.34}$$

where $C > 0$ is independent of problem data. Hence (3.26) follows from (3.34) and Lemma 3.3.

Remark. If (1.3) is replaced by

$$\theta_t - a(x,t)\theta_{xx} = f(x,t,u),$$

the existence of a solution can be proved by the same method, provided that $a \in C^1(\overline{\Omega_T})$ and f is continuous in its arguments and satisfies a growth condition such that

$$|f(x,t,u)| \le g(x,t) + C \cdot |u|.$$

By the aid of local estimates of the equation, the particular type of the boundary condition at $x = 1$ is not important for the method. Let us elaborate on this point a little more. Suppose a proper boundary condition for θ is given at $x = 1$. Then we have local estimates for θ in a neighborhood of $x = 1$. Once local estimates are established, it remains to give estimates on $(0, 1 - \epsilon) \times (0, T)$ for sufficiently small ϵ. This leads to the exact situation discussed in this paper once we consider the new function $\gamma = \theta\zeta$, where $\zeta = 0$ on $(1 - \epsilon, 1)$ and $\zeta = 1$ on $[0, 1 - 2\epsilon)$. This is also the viewpoint when we treat the case $0 < b < 1$ in the proof of Theorem 2.1.

Acknowledgment. The author would like to thank Professor John Cannon and Professor Yanping Lin for many helpful discussions and their generous supply of references.

REFERENCES

[1] J. R. Cannon, *The solution of the heat equation subject to the specification of energy*, Quart. Appl. Math., 21 (1963), pp. 155–160.

[2] J. R. Cannon, Y. Lin, and J. Van Der Hoek, *A quasi-linear parabolic equation with nonlocal boundary condition*, Rend. Mat. Appl. (7), 9 (1989), pp. 239–264.

[3] N. I. Ionkin, *Solution of a bounded value problem in heat condution with a nonclassical boundary condition*, Differential Equations, 13 (1977), pp. 294–304.

[4] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasi-linear Parabolic Eqations*, Transl. Math. Monographs, 23 (1968).

[5] V. L. Makarov and D. T. Kulyev, *Solution of a boundary value problem for a quasi-linear parabolic equation with a nonclassical boundary condition*, Differential Equations, 21 (1985), pp. 296–305.

[6] Y. Lin, *Parabolic partial differential equations subject to non-local boundary conditions*, Ph.D. thesis, Washington State University, Pullman, WA, 1988.

[7] J. L. Lions and E. Magenes, *Nonhomogenous Boundary Value Problems and Applications*, Springer-Verlag, New York, 1972.

[8] P. Shi and M. Shillor, *Design of contact patterns in one dimensional thermoelasticity*, in Theoretical Aspects of Industrial Design, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[9] R. S. Strichartz, *Multipliers on fractional Sobolev spaces*, J. Math. Mech., 16 (1967), pp. 1031–1060.

[10] K. Yosida, *Functional Analysis*, 6th ed., Spinger-Verlag, New York, 1980.

[11] N. I. Yurchuk, *Mixed problem with an integral condition for certain parobolic equations*, Differential Equations, 22 (1986), pp. 2117–2126.

SIAM J. MATH. ANAL.
Vol. 24, No. 1, pp. 59-75, January 1993

005

ON THE RIEMANN PROBLEM FOR A COMBUSTION MODEL*

TONG LI†

Abstract. This paper studies the Riemann problem for a combustion model and proves that the solution exists globally and the solution converges uniformly, away from the shock, to a travelling wave solution as $t \to +\infty$. The results are obtained by using the method of characteristics, a maximum principle and that the equation is a nonlinear conservation law.

Key words. detonation, conservation law, characteristic, maximum principle, shock wave, travelling wave, asymptotic behavior

AMS(MOS) subject classifications. 35L65, 35B40, 35B50, 76L05, 76J10

1. Introduction. We consider a combustion model

$$u_t + (\tfrac{1}{2}u^2 - q_0 z)_x = 0 \tag{1}$$

$$z_x = K\varphi(u)z. \tag{2}$$

introduced by Rosales and Majda [7]. The system can be considered as a simplified model for compressible Euler equations for reactive mixture if u is interpreted as a lumped variable with some features of pressure or temperature and z as the mass fraction of unburnt gas. In this interpretation, (1) and (2) describe a one-step irreversible chemical reaction, where $q_0 > 0$ is the amount of heat released by the reaction, K is the reaction rate, and $\varphi(u)$ has the ignition form

$$\varphi(u) = \begin{cases} 1 & u \geqq u_i \\ 0 & u < u_i \end{cases} \tag{3}$$

here the ignition temperature u_i satisfies $0 < u_i < \sqrt{2q_0}$.

The problem is well posed when the data are given by

$$u(x, 0) = u_0(x), \tag{4}$$

$$z(+\infty, t) = 1. \tag{5}$$

In this paper we consider the Riemann problem (1), (2), (5), and

$$u_0(x) = \begin{cases} u_0 & x \leqq 0, \\ 0 & x > 0. \end{cases} \tag{6}$$

Riemann problems are important because they are building blocks for solutions with general initial data and they also provide us a way to do numerical simulations.

One expects the solution of the Riemann problem to converge to a travelling wave (when it exists) as $t \to +\infty$. That is, the initial shock discontinuity triggers a detonation wave. In fact, the numerical results of Bourlioux [1] showed the convergence.

In this paper, we prove global existence of solutions to the Riemann problems and that the solutions converge to travelling wave solutions. The results are proved by using the method of characteristics, a maximum principle and the conservation property of (1). Independently, Levy obtained a similar result for a similar model in [4].

* Received by the editors September 17, 1991; accepted for publication (in revised form) May 5, 1992.

† Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, New York 10012. Current address: Institute for Advanced Study, School of Mathematics, Princeton, New Jersey 08540. This research was partially supported by National Science Foundation grant DMS 90-0226, U.S. Army grant DAAL 03-87-K-0063, and U.S. Air Force grant AFOSR-89-0203.

The main difficulty in the proof is that there is a shock discontinuity in the solution. One important aspect of the proof is to understand the evolution in time of the discontinuity.

The existence of the Riemann problem is established using a fixed point theorem based on an iteration procedure which is due to Ying and Teng [9]. Properties of the shock front $s(t)\in C^1[0,+\infty)$, namely that s' is bounded and s is convex, are obtained by employing a maximum principle and the method of characteristics. From these properties, we prove our a priori estimates for the solution which are needed for the convergence.

To prove that the solution of the Riemann problem converges to a shifted travelling wave, we use the conservation law (1) and the initial data to determine the shift. This technique was used before by Liu in the study of stability of shock waves [5]. It is used not only to identify the shift but also to obtain information which helps the convergence proof. The conservation laws and the comparison principle give us global L^1 control over the solution.

In § 2 we summarize results about travelling wave solutions for (1) and (2) by Rosales and Majda [7]. Section 3 is the proof of the global existence. Convergence is proved in § 4.

2. Travelling wave solutions. A travelling wave solution is a solution of the following form

$$(u(x,t),z(x,t))=(\psi(x-Dt),z(x-Dt)),$$

where D is the speed of the travelling wave solution. Let $\xi=x-Dt$. Plugging $(u(x,t), z(x,t))=(\psi(\xi),z(\xi))$ into (1) and (2), we have that $(\psi,z)(\xi)$ solves ordinary differential equations

$$-D\psi'+\psi\psi'=q_0z', \tag{7}$$

$$z'=K\varphi(\psi)z. \tag{8}$$

We impose boundary conditions

$$\lim_{\xi\to-\infty}(\psi,z)(\xi)=(u_0,0), \tag{9}$$

$$(\psi,z)(\xi)=(0,1),\qquad \xi\geqq 0. \tag{10}$$

Integrating (7) from $\xi<0$ to 0 and noticing that there is a shock discontinuity at $\xi=0$, we have

$$\psi(\xi)=D+(D^2-2q_0(1-z(\xi)))^{1/2},\qquad \xi<0. \tag{11}$$

Letting $\xi\to-\infty$ in the above equation and using (9), we get

$$D=D(u_0)=\frac{q_0}{u_0}+\frac{1}{2}u_0. \tag{12}$$

Integrating (8) and using (10), we have

$$z=e^{K\xi},\qquad \xi\leqq 0.$$

Inserting into (11), we get our travelling wave solution

$$(\psi(\xi),z(\xi))=\begin{cases}(0,1) & \xi\geqq 0\\ (D+(D^2-2q_0(1-e^{K\xi}))^{1/2},e^{K\xi}) & \xi<0.\end{cases}$$

It is easy to see that for the travelling wave to exist we have to require

$$u_0^2\geqq 2q_0. \tag{13}$$

We assume (13) throughout this paper. Corresponding to $u_0 = \sqrt{2q_0}$, we have $D = \sqrt{2q_0}$, which is the minimum travelling wave speed. The solution for the minimum value of D is called the Chapman-Jouguet (or CJ) detonation. A CJ detonation is characterized by being sonic with respect to the flow behind [2]. Because of this property, the CJ detonation plays a very special role in the stability theory. We will investigate this in a forthcoming paper.

From the explicit expression of the solution we see that the structure of a detonation wave is an ordinary nonreactive shock wave followed by a chemical reaction zone, which is the ZND (Zeldovich-Neumann-Doering) theory predicted.

3. Global existence. In this section we prove global existence and give global a priori estimates for solutions of the Riemann problem. We get the existence by means of a fixed point method based on an iteration procedure. A maximum principle and the method of characteristics enable us to get properties of the shock front. From these properties, we obtain our a priori estimates for the solution.

There is a shock discontinuity in the solution initiated from the Riemann initial data. Denote the shock wave position by $x = s(t)$. From the Rankine-Hugoniot condition and the initial condition, we have

$$\frac{ds}{dt} = \frac{1}{2} u(s(t), t), \tag{14}$$

$$s(0) = 0. \tag{15}$$

Writing (1) in characteristic form, we have

$$\frac{dx}{dt} = u(x, t),$$

$$\frac{du}{dt} = q_0 z_x.$$

Hence,

$$\frac{d^2x}{dt^2} = \frac{du}{dt} = q_0 z_x = q_0 K\varphi(u) z \geqq 0.$$

Therefore, every characteristic line $x = x(t)$ is convex in t and u increases along each characteristic line.

Let us simplify the problem by solving (2) in terms of the shock wave position.

For characteristic lines such that $x(t) \leqq s(t)$, it follows from the admissible conditions for shock waves [3] that $x(0) \leqq 0$. Furthermore, $u(x(t), t) \geqq u(x(0), 0) = u_0 > u_i$, $\varphi(u(x(t), t)) = 1$.

Similarly, for characteristic lines such that $x(t) > s(t)$, we have $x(0) > 0$. Furthermore, $u(x(t), t) = u(x(0), 0) = 0 < u_i$, $\varphi(u(x(t), t)) = 0$.

Plugging the value of φ into (2), we have that

$$z_x = \begin{cases} Kz & x \leqq s(t), \\ 0 & x > s(t). \end{cases}$$

Noticing that $\lim_{x \to +\infty} z(x, t) = 1$, z can be solved as

$$z(x, t) = \begin{cases} e^{K(x - s(t))} & x \leqq s(t), \\ 1 & x > s(t). \end{cases} \tag{16}$$

Plugging z into (1), the system is reduced to

$$u_t + uu_x = \begin{cases} q_0 K\, e^{K(x-s(t))} & x \leqq s(t), \\ 0 & x > s(t), \end{cases} \tag{17}$$

where $s(t)$ satisfies (14) and (15).

We now construct the solution by the iteration.

Consider first the following auxiliary problem.

$$u_t + uu_x = q_0 K\, e^{K(x-j(t))}, \tag{18}$$

$$u(x, 0) = u_0, \tag{19}$$

where $j \in E$, and

$$E = \{j : j \in C^1[0, T], j(0) = 0, j'(0) = \tfrac{1}{2}u_0, j'(t) \leqq D(u_0), j \text{ convex}\}.$$

Clearly, E is a closed bounded subset of $C^1[0, T]$.

We then iterate on j to find a fixed point of $s'(t) = \frac{1}{2}u(j(t), t)$ such that it is a solution of (14) and u is a solution of (17).

Lemma 3.1. *There exists a unique smooth solution u of problem* (18) *and* (19) *for all $t > 0$, which satisfies*

$$\frac{\partial u}{\partial x} \geqq 0, \tag{20}$$

$$\frac{d}{dt} u(j(t) + c, t) \geqq 0, \qquad c \leqq 0. \tag{21}$$

Proof. The classical solution of (18) and (19) can be constructed through its characteristic lines. Global solution exists if two characteristic lines never intersect.

Take any two characteristic lines $x_1(t)$ and $x_2(t)$ of $u(x, t)$, where $x_1(0) < x_2(0)$. We claim that the two characteristic lines $x_1(t)$ and $x_2(t)$ never intersect.

To prove the claim, suppose for the contrary that at time $t = t_0 > 0$ they intersect. We have $x_1(t_0) = x_2(t_0)$ and $x_1(t) < x_2(t)$, $0 \leqq t < t_0$. Hence, $d^2x_1/dt^2 < d^2x_2/dt^2$, $0 \leqq t < t_0$.

By Taylor expansion with remainder, there is some $\xi \in [0, t_0]$ such that

$$\begin{aligned} 0 &= x_1(t_0) - x_2(t_0) \\ &= (x_1 - x_2)(0) + \frac{d(x_1 - x_2)(0)}{dt} t_0 + \frac{1}{2} \frac{d^2(x_1 - x_2)(\xi)}{dt^2} t_0^2 \\ &\leqq (x_1 - x_2)(0) + (u(x_1(0), 0) - u(x_2(0), 0)) t_0 \\ &= (x_1 - x_2)(0) \\ &< 0, \end{aligned}$$

which is a contradiction. Thus $x_1(t)$ and $x_2(t)$ never intersect.

Given any two points (x_{10}, t_0) and (x_{20}, t_0), $t_0 > 0$ and $x_{10} < x_{20}$, draw characteristic lines $x_1(t)$ and $x_2(t)$ backwards in time. Since any two characteristic lines never intersect, we have

$$x_1(t) < x_2(t), \qquad 0 \leqq t \leqq t_0.$$

Integrating (18) along the characteristic lines, we have

$$u(x_1(t_0), t_0) < u(x_2(t_0), t_0).$$

Noticing that u is smooth, we have

$$\frac{\partial u}{\partial x} \geqq 0 \quad \text{for } t>0.$$

This proves (20).

To prove (21) we need a transformation

$$r=t,$$
$$y=x-j(t).$$

Then (18) becomes

$$u_r+(u-j')u_y=q_0K\,e^{Ky}.$$

Let $p=\partial u/\partial r$. Differentiating the above equation with respect to r, we have

$$\frac{dp}{dr}+\frac{\partial u}{\partial y}p=\frac{\partial u}{\partial y}j'',$$
$$p(y,0)=q_0K\,e^{Ky}>0,$$

where $d/dr=(\partial/\partial r)+(u-j')\,(\partial/\partial y)$ is derivative in the characteristic direction.

Assuming that j is smooth for the moment, we have $j''\geqq 0$. From (20), $\partial u/\partial y=\partial u/\partial x\geqq 0$. Now applying a maximum principle in the above equations for p, we have that $p(y,r)\geqq 0$ for any $r\geqq 0$ and any y. That is

$$p=\frac{\partial u}{\partial r}=\frac{\partial u}{\partial t}+j'\frac{\partial u}{\partial x}\geqq 0.$$

In particular, this holds at $y=c$, $c\leqq 0$, i.e., at $x=j(t)+c$, we have

$$\frac{\partial}{\partial t}u(j(t)+c,t)+j'(t)\frac{\partial}{\partial x}u(j(t)+c,t)\geqq 0.$$

That is,

$$\frac{d}{dt}u(j(t)+c,t)\geqq 0.$$

If j is not smooth enough, we arrive at our conclusion by approximating j by smooth functions and passing to the limit, since the result only involves the first derivative of j. □

We next prove a comparison principle for solutions of (1) and (2), which will be useful for qualitative study of the solutions.

In the following theorem the initial value has the following form:

$$u(x,0)=\begin{cases} u_0(x) & x\leqq s(0), \\ 0 & x>s(0), \end{cases}$$

where $u_0(x)\geqq 0$ is a nondecreasing function and $s(0)$ is the initial shock wave position.

THEOREM 3.2 (A comparison principle). *Suppose that $u_1(x,t)$ and $u_2(x,t)$ are solutions of* (1) *and* (2) *with nondecreasing initial data $u_{10}(x)$ and $u_{20}(x)$ and shock wave positions $s_1(t)$ and $s_2(t)$, respectively. If*

$$s_1(0)<s_2(0)$$

and

$$u_{10}(x)\geqq u_{20}(x), \qquad x\leqq s_1(0),$$

then there is some $T>0$ such that for $0<t<T$, we have

$$u_1(x,t)>u_2(x,t), \qquad x\leqq s_1(t).$$

Proof. Since $s_1(0) < s_2(0)$, there exists a $T > 0$ such that

$$s_1(t) < s_2(t), \qquad 0 < t < T.$$

From point (x, t_0), $0 < t_0 < T$, draw characteristic lines $x_1(t)$ and $x_2(t)$ of u_1 and u_2 backwards, respectively.

Step 1. Suppose that (x, t_0) is the first intersection point of the two characteristic lines. Then we claim $u_1(x, t_0) > u_2(x, t_0)$.

Suppose for the contrary that $u_1(x, t_0) \leqq u_2(x, t_0)$. Then, $x_1(t) > x_2(t)$, for $0 \leqq t < t_0$. We also know that $s_1(t) < s_2(t)$, for $0 \leqq t < t_0$. So, $x_1(t) - s_1(t) > x_2(t) - s_2(t)$. Integrating the equations for u_1 and u_2 along their characteristic lines $x_1(t)$ and $x_2(t)$, respectively, we have that

$$u_1(x, t_0) > u_2(x, t_0),$$

which contradicts our assumption that $u_1(x, t_0) \leqq u_2(x, t_0)$. Thus, $u_1(x, t_0) > u_2(x, t_0)$.

Step 2. We claim that any two characteristic lines of u_1 and u_2 intersect at most once. If two characteristic lines intersect at $t = t_0$, by Step 1, we have

$$\frac{dx_1}{dt} = u_1(x, t_0) > u_2(x, t_0) = \frac{dx_2}{dt},$$

$$x_1(t_0) = x_2(t_0).$$

We thus have

$$x_1(t) > x_2(t), \qquad t_0 < t < t_0 + \delta$$

for some $\delta > 0$.

Claim. $x_1(t) > x_2(t)$ for all $t > t_0$.

If this is not true, then there is a $t_1 > t_0$ such that $x_1(t_1) = x_2(t_1)$ and $x_1(t) > x_2(t)$, $t_0 < t < t_1$. It follows that $d^2x_1/dt^2 > d^2x_2/dt^2$, $t_0 < t < t_1$. By Taylor expansion with remainder, there is $t_0 < \xi < t_1$ such that

$$\begin{aligned} 0 &= x_1(t_1) - x_2(t_1) \\ &= x_1(t_0) - x_2(t_0) + \frac{d(x_1 - x_2)(t_0)}{dt}(t_1 - t_0) \\ &\quad + \frac{1}{2}\frac{d^2(x_1 - x_2)(\xi)}{dt^2}(t_1 - t_0)^2 > 0, \end{aligned}$$

a contradiction. Hence, no two characteristic lines of u_1 and u_2 intersect more than once.

By Step 1, we have

$$u_1(x, t_0) > u_2(x, t_0)$$

for all $x \leqq s_1(t_0)$, and $0 < t_0 < T$. □

Using the comparison principle, we get upper bound on the solution in the following lemma.

LEMMA 3.3.

$$\tfrac{1}{2}u(j(t), t) < D(u_0), \qquad t > 0. \tag{22}$$

Proof. From point $(j(t_0), t_0)$, draw a straight line with slope $D(u_0)$. Since $j'(t) \leqq D(u_0)$ (by the choice of j), we have

$$x(t) = j(t_0) + D(u_0)(t - t_0) < j(t) \quad \text{for } 0 \leqq t < t_0.$$

$x(t)=j(t_0)+D(u_0)(t-t_0)$ can be viewed as the shock wave position of the travelling wave $\psi(x-Dt+c)$, where $c=-j(t_0)+Dt_0>0$. Comparing initial data of $\psi(x-Dt+c)$ and $u(x,t)$, we have

$$\psi(x+c)\geqq u(x,0), \qquad x+c\leqq 0.$$

Applying the comparison principle in Theorem 3.2 to $\psi(x-Dt+c)$ and $u(x,t)$, we get

$$\psi(x-Dt+c)>u(x,t), \qquad x-Dt+c\leqq 0 \quad \text{for } 0\leqq t\leqq t_0.$$

In particular, at point $(j(t_0), t_0)$,

$$2D(u_0)=\psi(0)>u(j(t_0), t_0).$$

That proves the lemma. □

Let s be the solution of

$$s(0)=0,$$
$$s'(t)=\tfrac{1}{2}u(j(t), t).$$

By Lemma 3.3,

$$s'(t)<D(u_0).$$

By Lemma 3.1,

$$s''(t)=\frac{1}{2}\frac{d}{dt}u(j(t), t)\geqq 0.$$

Hence, $s\in E$ and is twice differentiable.

Define an operator $A: E\to C^1[0, T]$

$$s=Aj.$$

Clearly, $AE\subset E$, i.e., A maps E into itself.

LEMMA 3.4. *The mapping A has a fixed point.*

Proof. By the Arzela-Ascoli theorem, it is enough to prove that $s''=(Aj)''$ is bounded for all $j\in E$ (since then AE will be a bounded subset of $C^2[0, T]$, and hence, a compact subset of $E\subset C^1[0, T]$). This will assure the compactness necessary for the Schauder fixed point theorem.

By Lemma 3.3,

$$0\leqq j'(t)\leqq D.$$

A straightforward calculation from conditions (12) and (13) yields

$$u_0\geqq D.$$

Finally, using $u_x\geqq 0$ from Lemma 3.1, we have

$$\begin{aligned}0\leqq s''&=\frac{1}{2}\frac{d}{dt}u(j(t), t)\\&=\frac{1}{2}(u_t+j'(t)u_x)\\&=\frac{1}{2}(u_t+uu_x)+\frac{1}{2}(j'(t)-u)u_x\\&\leqq\frac{1}{2}q_0K+\frac{1}{2}(D-u_0)u_x\\&\leqq\frac{1}{2}q_0K,\end{aligned}$$

which is bounded uniformly with respect to $j \in E$. Hence, by the Schauder fixed point theorem A has a fixed point. □

Using this fixed point, we construct the following solution of Riemann problem (1), (2), (6), and (5):

$$u(x, t) = \begin{cases} u_j(x, t) & x \leqq s(t), \\ 0 & x > s(t), \end{cases} \tag{23}$$

$$z(x, t) = \begin{cases} e^{K(x - s(t))} & x \leqq s(t), \\ 1 & x > s(t), \end{cases} \tag{24}$$

where u_j is the solution of (18) and (19).

It is easy to check that (u, z) is a solution.

Next, we prove that the above-defined solution is unique. This allows us to extend the solution to $t = +\infty$.

LEMMA 3.5. *Solutions of form* (23) *and* (24) *are unique.*

Proof. Suppose that there are two solutions $u(x, t)$ and $v(x, t)$ with shock wave positions $s_1(t)$ and $s_2(t)$, respectively.

Because of the uniqueness of solutions of (18) and (19), we need only to show that $s_1(t) = s_2(t)$.

Translate s_2 along the x axis to the right of s_1 until they just touch. If the lowest touch point is $(s_1(t_0), t_0)$, then $t_0 = 0$.

Suppose for the contrary that $t_0 > 0$. Let $s_2 + \delta$ be the translation of s_2 such that $(s_1(t_0), t_0)$ is on $s_2 + \delta$. Then $\delta > 0$ and $s_1(t) < s_2(t) + \delta$ for $0 < t < t_0$.

Let

$$v_1(x, t) = v(x - \delta, t),$$
$$z_1(x, t) = z(x - \delta, t),$$

then (v_1, z_1) satisfies

$$v_{1t} + v_1 v_{1x} = q_0 z_{1x},$$
$$z_{1x} = q_0 \phi(v_1) z_1,$$
$$v_1(x, 0) = \begin{cases} u_0 & x \leqq \delta, \\ 0 & x > \delta, \end{cases}$$
$$z_1(x, 0) = 1.$$

In fact,

$$z_1(x, t) = \begin{cases} e^{K(x - (s_2(t) + \delta))} & x \leqq s_2(t) + \delta, \\ 1 & x > s_2(t) + \delta. \end{cases}$$

At time $t = 0$, we have

$$\int_{-\infty}^{+\infty} (u - v_1)(x, 0)\, dx = -\delta u_0 < 0. \tag{25}$$

It is easy to check that the conditions for the comparison principle hold. Applying the comparison principle in Theorem 3.2 to u and v_1, we have $u(x, t_0) > v_1(x, t_0)$ for $x \leqq s_1(t_0) = s_2(t_0) + \delta$. But this implies

$$\int_{-\infty}^{+\infty} (u - v_1)(x, t_0)\, dx = \int_{-\infty}^{s_1(t_0)} (u - v_1)(x, t_0)\, dx > 0,$$

which contradicts (25) since by the conservation law (1)

$$\int_{-\infty}^{+\infty} (u - v_1)(x, t)\, dx$$

is a constant independent of t.

Hence, $t_0 = 0$, which implies that $\delta = 0$. It follows that $s_1(t) \leqq s_2(t)$. Similarly, $s_2(t) \leqq s_1(t)$. Hence $s_2(t) = s_1(t)$. □

A nice property of the solution is given in the next lemma.

LEMMA 3.6. *The solution of the Riemann problem is convex in x, that is,*

$$\frac{\partial^2 u}{\partial x^2}(x, t) \geqq 0 \quad \textit{for } x \leqq s(t).$$

Proof. Differentiating equation (1) with respect to x twice, we have

$$u_{xxt} + u u_{xxx} + 3 u_x u_{xx} = q_0 z_{xxx}.$$

Rewriting the above equation in the characteristic form, we have

$$\frac{du_{xx}}{dt} + 3 u_x u_{xx} = q_0 K^3\, e^{K(x - s(t))} > 0.$$

Initially,

$$u_{xx}(x, 0) = 0 \quad \text{for } x \leqq 0.$$

Using a maximum principle along the characteristic direction, we have

$$u_{xx}(x, t) \geqq 0 \quad \text{for } x \leqq s(t). \qquad \square$$

Now we have the following theorem.

THEOREM 3.7. *Solution of* (1), (2), (6), *and* (5) *exists globally and satisfies*

$$0 \leqq u \leqq M, \quad 0 \leqq u_x \leqq M, \quad |u_t| \leqq M, \quad u_{xx} \geqq 0,$$

$$0 \leqq z \leqq 1, \quad 0 \leqq z_x \leqq K, \quad |z_t| \leqq M,$$

where M depends only on q_0, K, u_0. All the estimates involving derivatives are valid away from the shock curve $(s(t), t)$.

Furthermore, $s(t)$ satisfies

$$s'' \geqq 0$$

and

$$\lim_{t \to +\infty} s'(t)$$

exists.

Proof. From Lemmas 3.1 and 3.5, $u_j(x, t)$ and $s = s(t)$ are defined for all $t > 0$. Therefore, the solution of (1), (2), (6), and (5) exists for all $t > 0$ by the construction (23) and (24).

It follows from (20), $u_x \geqq 0$, that $u(x, t) \leqq u(s(t), t)$ for $x \leqq s(t)$. Furthermore, $u(s(t), t) \leqq 2D$ by Lemma 3.3. Hence, u is bounded.

Let $a = u_x$. Differentiating equation (1) with respect to x and writing the resulting equation in characteristic form, we get

$$\frac{da}{dt} + a^2 = q_0 K^2\, e^{K(x - s(t))}.$$

Integrating along the characteristic line $x=x(t)$ from 0 to t, we obtain

$$0 \leqq a(x(t),t) \leqq a(x(0),0)+\int_0^t q_0K^2\, e^{K(x(t)-s(t))}\, dt$$
$$=K(u(x(t),t)-u(x(0),0)) \leqq M.$$

Using (17),

$$u_t=q_0K\, e^{K(x-s(t))}-uu_x,$$

we can get the desired bound for u_t. From (24), clearly, z and z_x satisfy the above estimates. As for z_t, we have

$$z_t(x,t)=-Ks'(t)\, e^{K(x-s(t))}, \qquad x<s(t)$$

which is bounded.

By definition of s and Lemma 3.1, we have

$$s'(t)=\tfrac{1}{2}u(s(t),t) \leqq D(u_0),$$
$$s''(t) \geqq 0.$$

Hence, $\lim_{t\to+\infty} s'(t)$ exists. □

4. Convergence to travelling waves. In this section we prove that the solution of the Riemann problem (1), (2), (6), and (5) converges to a shifted travelling wave. The shift is determined from the initial data. This technique was used before by Liu [5]. We use it not only for identifying the shift but also for showing the convergence. The conservation laws and the comparison principle give us global L^1 control over the solution.

Let

$$x_0=-\frac{1}{u_0}\int_{-\infty}^{+\infty}\{u(x,0)-\psi(x)\}\, dx. \tag{26}$$

From the initial data, $x_0>0$.

Our final goal is to show that the solution $(u,z)\,(x,t)$ converges to $(\psi,z)\,(x-Dt+x_0)$ as $t\to+\infty$. This means that the final wave front is determined by its initial total mass. We note that the mass difference between any two solutions is a conserved quantity.

LEMMA 4.1. *If x_0 is determined as above, then*

$$\int_{-\infty}^{+\infty}(u(x,0)-\psi(x+x_0))\, dx=0. \tag{27}$$

Proof. First of all,

$$0>\int_{-\infty}^{+\infty}(u(x,0)-\psi(x))\, dx$$
$$=\int_{-\infty}^{0}(u_0-\psi(x))\, dx$$
$$=-\int_{-\infty}^{0}\frac{2q_0\, e^{Kx}}{(D^2-2q_0)^{1/2}+(D^2-2q_0(1-e^{Kx}))^{1/2}}\, dx$$
$$\geqq -\int_{-\infty}^{0}(2q_0\, e^{Kx})^{1/2}\, dx>-\infty.$$

Thus $u(x,0)-\psi(x)$ is integrable and so is $\psi(x+x_0)-\psi(x)$.

$$\begin{aligned}\int_{-\infty}^{+\infty}(\psi(x+x_0)-\psi(x))\,dx&=\left(\int_{-\infty}^{-x_0}+\int_{-x_0}^{+\infty}\right)(\psi(x+x_0)-\psi(x))\,dx\\&=\int_{-\infty}^{-x_0}(\psi(x+x_0)-u_0)+(u_0-\psi(x))\,dx+\int_{-x_0}^{0}[(u_0-\psi(x))-u_0]\,dx\\&=\int_{-\infty}^{-x_0}(\psi(x+x_0)-u_0)\,dx+\int_{-\infty}^{0}(u_0-\psi(x))\,dx-x_0u_0\\&=\int_{-\infty}^{0}(\psi(x)-u_0)\,dx+\int_{-\infty}^{0}(u_0-\psi(x))\,dx-x_0u_0\\&=-x_0u_0.\end{aligned}$$

The conclusion then follows easily from (26). □

For the travelling wave $(\psi, z)\ (x-Dt+x_0)$, its equations in characteristic form are

$$\frac{dx_1}{dt}=\psi,$$

$$\frac{d\psi}{dt}=q_0K\,e^{K(x_1(t)-s_1(t))},$$

where $s_1(t)=Dt-x_0$ is the shock wave position for the travelling wave. In the next lemma, we use the conservation law to estimate the lower bound of $s(t)$.

LEMMA 4.2. *Let $s_1(t)=Dt-x_0$ and $s(t)$ be the shock wave positions for $\psi(x-Dt+x_0)$ and $u(x,t)$, respectively.*

Then $s_1(t)<s(t)$ for $t>0$.

Proof. First, since $s_1(0)=-x_0<0=s(0)$, we have $s_1(t)<s(t)$ for $t<\delta$, where $\delta>0$.

Let $T=\sup\{t\,|\,s_1(\tau)<s(\tau)\text{for }\tau<t\}$. We want to prove $T=+\infty$. If this is not true, we have that $T=t_0<+\infty$. Hence

$$s_1(t_0)=s(t_0),$$
$$s_1(t)<s(t)\quad\text{for } t<t_0.$$

Applying the comparison principle Theorem 3.2 to ψ and u, we see that

$$\psi(x-Dt_0+x_0)>u(x,t_0)\quad\text{for } x\leqq s_1(t_0)=s(t_0),$$

and hence

$$\int_{-\infty}^{+\infty}[\psi(x-Dt_0+x_0)-u(x,t_0)]\,dx=\int_{-\infty}^{s_1(t_0)}[\psi(x-Dt_0+x_0)-u(x,t_0)]\,dx>0.$$

Using the conservation property of (1) and also the identity (27), we have

$$0<\int_{-\infty}^{\infty}[\psi(x-Dt_0+x_0)-u(x,t_0)]\,dx=\int_{-\infty}^{+\infty}[\psi(x+x_0)-u(x,0)]\,dx=0,$$

a contradiction. Therefore, $s_1(t)<s(t)$ for all $t>0$. □

LEMMA 4.3.

$$0<\psi(x-Dt+x_0)-u(x,t)<C\,e^{K(x-Dt+x_0)/2}$$

for $x-Dt+x_0<0$.

Proof. For $x-Dt+x_0<0$, we have $x<s_1(t)=Dt-x_0<s(t)$ by Lemma 4.2. Using the comparison principle Theorem 3.2, we have

$$\psi(x-Dt+x_0)>u(x,t)>u_0\quad\text{for } x<s_1(t).$$

Therefore,

$$\begin{aligned}0&<\psi(x-Dt+x_0)-u(x,t)\\&<\psi(x-Dt+x_0)-u_0\\&=\frac{2q_0\,e^{K(x-Dt+x_0)}}{(D^2-2q_0)^{1/2}+(D^2-2q_0(1-e^{K(x-Dt+x_0)}))^{1/2}}\\&\leqq\sqrt{2q_0}\,e^{K(x-Dt+x_0)/2}.\end{aligned}\qquad\square$$

This lemma implies that the difference between u and the travelling wave is integrable.

Next, we prove that $x=s(t)$ has an asymptotic line $x=Dt-\delta$ as $t\to+\infty$. Eventually we will show $\delta=x_0$.

LEMMA 4.4.

$$\lim_{t\to+\infty}(s(t)-Dt+\delta)=0,$$

$$\lim_{t\to+\infty}s'(t)=D(u_0)$$

for some δ, $0<\delta\leqq x_0$.

Proof. Let $E(t)=s(t)-Dt+x_0$. Then $E(t)>0$ and $E'(t)=s'(t)-D\leqq 0$ by Lemmas 4.2 and 3.3. Therefore, $\lim_{t\to+\infty}E(t)$ exists. Hence, there is some constant δ, with $0<\delta\leqq x_0$, such that

$$\lim_{t\to+\infty}s(t)-Dt+\delta=0.$$

By Theorem 3.7, $\lim_{t\to+\infty}s'(t)$ exists. Therefore,

$$\lim_{t\to+\infty}s'(t)=D.\qquad\square$$

Next, we prove that the solution to the Riemann problem are asymptotically functions of one variable $x-Dt+\delta$.

LEMMA 4.5. *There are Lipschitz functions* $u_\infty(\xi)$ *and* $z_\infty(\xi)$ *such that*

$$\lim_{t\to+\infty}\ \sup_{x-Dt+\delta\leqq 0}|(u,z)(x,t)-(u_\infty,z_\infty)(x-Dt+\delta)|=0.$$

Proof. Let

$$\xi=x-Dt+\delta,$$

$$u(x,t)=U(\xi,t),$$

$$z(x,t)=Z(\xi,t).$$

Let $E(t)=s(t)-Dt+\delta$. By Lemma 4.4, $E(t)$ decreases to zero and $s'(t)$ increases to D as $t\to+\infty$.

For any $\varepsilon>0$, there is a $T>0$, such that if t, t_1, $t_2>T$,

$$0<E(t)<\varepsilon,\qquad(28)$$

$$2|s'(t_1)-s'(t_2)|<\varepsilon,$$

$$1-\varepsilon<e^{-KE(t)}<1.\qquad(29)$$

Since $E(t)>0\geqq\xi$, we have $s(t)>x$. By (24),

$$Z(\xi,t)=z(x,t)=e^{K(x-s(t))}=e^{K\xi}\,e^{-KE(t)}.$$

Using (29),

$$(1-\varepsilon)\, e^{K\xi} < Z(\xi, t) < e^{K\xi}, \qquad \xi \leqq 0.$$

Letting

$$z_\infty(\xi) = e^{K\xi}, \qquad \xi \leqq 0,$$

we have

$$\lim_{t\to+\infty} \sup_{\xi\leqq 0} |Z(\xi, t) - z_\infty(\xi)| = 0.$$

We now look at the characteristic lines of $U(\xi, t)$ given by

$$\frac{d\xi}{dt} = \frac{dx}{dt} - D = U - D$$

and

$$\frac{d^2\xi}{dt^2} = \frac{d^2x}{dt^2} = \frac{du}{dt} = q_0 z_x = q_0 K\, e^{K\xi}\, e^{-KE(t)}.$$

Multiplying both sides of the above equality by $d\xi/dt$ and using inequality (29), we have

$$-\varepsilon q_0 K\, e^{K\xi} \frac{d\xi}{dt} < \frac{d\xi}{dt}\frac{d^2\xi}{dt^2} - q_0 K\, e^{K\xi}\frac{d\xi}{dt} < 0.$$

Integrating the above inequalities from t_1 to t_2 and using (29), we see that

$$-\varepsilon q_0\, e^{K\xi}\Big|_{t_1}^{t_2} < \frac{1}{2}\left(\frac{d\xi}{dt}\right)^2 - q_0\, e^{K\xi}\Big|_{t_1}^{t_2} < 0.$$

Noticing that $d\xi/dt = U - D$, we have

$$|\tfrac{1}{2}(U(\xi(t), t) - D)^2 - q_0\, e^{K\xi(t)}|\Big|_{t_1}^{t_2} < q_0\varepsilon. \tag{30}$$

Draw characteristic lines $\xi_1(t)$ and $\xi_2(t)$ from (ξ, t_1) and (ξ, t_2), respectively. Let t_1^* and t_2^* be the time at which $\xi_1(t)$ and $\xi_2(t)$ intersect $s(t)$, respectively. It follows from (28) that

$$\begin{aligned} &|U(\xi_2(t_2^*), t_2^*) - U(\xi_1(t_1^*), t_1^*)| \\ &\quad = |u(s(t_2^*), t_2^*) - u(s(t_1^*), t_1^*)| \\ &\quad = 2|s'(t_2^*) - s'(t_1^*)| < \varepsilon. \end{aligned} \tag{31}$$

We then estimate

$$\begin{aligned} &|\tfrac{1}{2}(U(\xi, t) - D)^2|_{t_1}^{t_2}| \\ &\quad \leqq |\tfrac{1}{2}(U(\xi_2(t_2), t_2) - D)^2 - q_0\, e^{K\xi_2(t_2)} - \tfrac{1}{2}(U(\xi_2(t_2^*), t_2^*) - D)^2 + q_0\, e^{K\xi_2(t_2^*)}| \\ &\qquad + |\tfrac{1}{2}(U(\xi_1(t_1), t_1) - D)^2 - q_0\, e^{K\xi_1(t_1)} - \tfrac{1}{2}(U(\xi_1(t_1^*), t_1^*) - D)^2 + q_0\, e^{K\xi_1(t_1^*)}| \\ &\qquad + |q_0\, e^{KE(t_1^*)} - q_0\, e^{KE(t_2^*)}| \\ &\qquad + |\tfrac{1}{2}(U(\xi_2(t_2^*), t_2^*) - D)^2 - \tfrac{1}{2}(U(\xi_1(t_1^*), t_1^*) - D)^2| \\ &\quad < c\varepsilon, \end{aligned}$$

where the last inequality follows from (30), (31), and (29).

Therefore, $\lim_{t\to+\infty}\frac{1}{2}(U(\xi,t)-D)^2$ exists and so does $\lim_{t\to+\infty}U(\xi,t)$. Denote this limit by $u_\infty(\xi)$. By Theorem 3.7, u, z, and their first derivatives are bounded. Hence $u_\infty(\xi)$, $z_\infty(\xi)$ are Lipschitz, and

$$\lim_{t\to+\infty}\sup_{\xi\leqq 0}|U(\xi,t)-u_\infty(\xi)|=0. \qquad \square$$

LEMMA 4.6.

$$\lim_{t\to+\infty}(U,Z)(\xi,t)=(0,1)=(u_\infty(\xi),z_\infty(\xi)),$$

where $\xi=x-Dt+\delta>0$.

Proof. If $\xi>0$, i.e., $x-Dt+\delta>0$, there is a $T>0$, such that for $t>T$,

$$0<E(t)=s(t)-Dt+\delta<x-Dt+\delta.$$

That is,

$$x-s(t)>0.$$

From (24) and (23), we have that

$$Z(\xi,t)=z(x,t)=1, \qquad U(\xi,t)=u(x,t)=0,$$

provided $t>T$.

Taking

$$(u_\infty,z_\infty)(\xi)=(0,1)$$

where $\xi>0$, the lemma is proved. $\square$

LEMMA 4.7. $(u_\infty,z_\infty)(\xi)$ *is a travelling wave solution* $(\psi,z)(\xi)$ *of* (1) *and* (2), *i.e.*,

$$(u_\infty,z_\infty)(\xi)=(\psi,z)(\xi), \quad \textit{where } \xi=x-Dt+\delta.$$

Proof. (u,z) is a weak solution of (1) and (2) if and only if

$$\iint \phi_t(x,t)u(x,t)+\phi_x(x,t)\left[\frac{1}{2}u^2(x,t)-q_0z(x,t)\right]dx\,dt=0$$

for any smooth function ϕ with compact support in $\{(x,t)|t>0\}$.

Taking $\phi(x-DT,t-T)$ as the test function in the above equation, then

$$\iint \phi_t(x-DT,t-T)u(x,t)+\phi_x(x-DT,t-T)\left[\frac{1}{2}u^2(x,t)-q_0z(x,t)\right]dx\,dt=0.$$

A change of variables in the above integral gives

$$\iint \phi_t(x,t)u(x+DT,t+T)$$
$$+\phi_x(x,t)\left[\frac{1}{2}u^2(x+DT,t+T)-q_0z(x+DT,t+T)\right]dx\,dt=0.$$

Hence

$$\iint \phi_t(x,t)U(x-Dt+\delta,t+T)$$
$$+\phi_x(x,t)\left[\frac{1}{2}U^2(x-Dt+\delta,t+T)-q_0Z(x-Dt+\delta,t+T)\right]dx\,dt=0.$$

Letting $T \to +\infty$,

$$\iint \phi_t(x,t)u_\infty(x-Dt+\delta)+\phi_x(x,t)\left[\frac{1}{2}u_\infty^2(x-Dt+\delta)-q_0z_\infty(x-Dt+\delta)\right]dx\,dt=0$$

for all ϕ.

Using Lemmas 4.5 and 4.6,

$$z_{\infty x}(x-Dt+\delta)=\begin{cases}Kz_\infty & x-Dt+\delta\leqq 0\\ 1 & x-Dt+\delta>0\end{cases}=K\varphi(u_\infty)z_\infty.$$

Thus, $(u_\infty, z_\infty)(x-Dt+\delta)$ is indeed a weak solution of (1) and (2).

We now show that $(u_\infty, z_\infty)(x-Dt+\delta)$ satisfies (7)-(10), which are equations for the travelling wave solution.

Clearly, $z_\infty(\xi)$ satisfies (8) and is smooth for $\xi\leqq 0$. By Lemma 4.5 $(u_\infty, z_\infty)(\xi)$ is Lipschitz for $\xi\leqq 0$, and hence absolutely continuous. Furthermore, (u_∞, z_∞) satisfies (7) almost everywhere. Using Lemma 4.6, we get

$$(u_\infty, z_\infty)(x-Dt+\delta)=(0,1)=(\psi,z)(x-Dt+\delta),\qquad x-Dt+\delta>0.$$

That is, (u_∞, z_∞) satisfies (10).

By Lemma 4.2,

$$u_0<u(x,t)<\psi(x-Dt+x_0),\qquad x-Dt+\delta<0.$$

Noticing that $\psi(-\infty)=u_0$, we get

$$\lim_{x-Dt+\delta\to-\infty} u_\infty(x-Dt+\delta)=u_0.$$

From Lemma 4.5,

$$\lim_{x-Dt+\delta\to-\infty} z_\infty(x-Dt+\delta)=0.$$

Hence (9) is satisfied.

Using the uniqueness of solution of (7)-(10),

$$(u_\infty, z_\infty)(x-Dt+\delta)=(\psi,z)(x-Dt+\delta),\qquad 0<\delta\leqq x_0.$$ □

LEMMA 4.8.

$$\delta=x_0.$$

That is, $(u, z)(x, t)$ *converges to the travelling wave* $(\psi, z)(x-Dt+x_0)$.

Proof. From Lemma 4.1, we have

$$\int_{-\infty}^{+\infty}[\psi(x+x_0)-u(x,0)]\,dx=0.$$

Using the conservation law and Lemma 4.3, we have that for any $t>0$,

$$\int_{-\infty}^{+\infty}(\psi(x-Dt+x_0)-u(x,t))\,dx=0.$$

Hence,

$$\int_{-\infty}^{+\infty}(\psi(x-Dt+x_0)-\psi(x-Dt+\delta))\,dx=0.$$

That is, $\delta=x_0$. □

We summarize our results in the following theorem.

THEOREM 4.9. *The solution of* (1), (2), (6), *and* (5) *converges uniformly to the travelling wave solution and the shock front is asymptotically linear*:

$$\lim_{t\to+\infty}\ \sup_{x-Dt+x_0\leqq 0}|(u,z)(x,t)-(\psi,z)(x-Dt+x_0)|=0,$$

$$\lim_{t\to+\infty}(u,z)(x,t)=(0,1)\quad \textit{for } x-Dt+x_0>0$$

and

$$\lim_{t\to+\infty}(s(t)-Dt+x_0)=0,$$

where x_0 *satisfies*

$$\int_{-\infty}^{\infty}[u(x,0)-\psi(x)]\,dx=-u_0x_0.$$

We also have the convergence result in L^p norm ($p\geqq 1$). This is a consequence of the conservation law and the above result.

COROLLARY 4.10.

$$\lim_{t\to+\infty}|(u,z)(x,t)-(\psi,z)(x-Dt+x_0)|_p=0,\qquad p\geqq 1.$$

Proof. We only prove the results for $p=1$. Results for $p>1$ are easy consequences of $p=1$ and Theorem 4.9.

Let $s_1(t)=Dt-x_0$. By Lemma 4.3 and the conservation law,

$$\int_{-\infty}^{s_1(t)}[\psi(x-Dt+x_0)-u(x,t)]\,dx=-\int_{s_1(t)}^{s(t)}[\psi(x-Dt+x_0)-u(x,t)]\,dx.$$

Therefore,

$$\begin{aligned}&|u(x,t)-\psi(x-Dt+x_0)|_1\\&=\int_{-\infty}^{s_1(t)}[\psi(x-Dt+x_0)-u(x,t)]\,dx+\int_{s_1(t)}^{s(t)}[u(x,t)-\psi(x-Dt+x_0)]\,dx\\&=2\int_{s_1(t)}^{s(t)}[u(x,t)-\psi(x-Dt+x_0)]\,dx\\&\leqq 2(s(t)-s_1(t))2D\to 0,\qquad t\to+\infty.\end{aligned}$$

A similar result for z directly follows from $s(t)-s_1(t)\to 0$. □

Acknowledgments. This is a part of my thesis. It is a pleasure to express my gratitude to my thesis advisor, Professor Tai-Ping Liu, for all his guidance and encouragement. Thanks are also due to A. Bourlioux for showing me her numerical results.

REFERENCES

[1] A. BOURLIOUX, *Analysis of numerical methods for a simplified detonation model*, Ph.D thesis, Princeton University, Princeton, NJ, 1991.

[2] W. FICKETT AND W. C. DAVIS, *Detonation*, University of California Press, Berkeley, CA, 1979.

[3] P. D. LAX, *Hyperbolic systems of conservation laws*, II, Comm. Pure Appl. Math., 10 (1957), pp. 537-566.

[4] A. LEVY, *On Majda's model for dynamic combustion*, Comm. Partial Differential Equations, 17 (1992), pp. 657-698.

[5] T. P. Liu, *Nonlinear stability of shock waves for viscous conservation laws*, Mem. Amer. Math. Soc. 328, American Mathematical Society, Providence, RI, 1985.

[6] A. Majda, *High Mach number combustion*, Lectures in Appl. Math. 24, American Mathematical Society, Providence, RI, 1986.

[7] R. Rosales and A. Majda, *Weakly nonlinear detonation waves*, SIAM J. Appl. Math., 43 (1983), pp. 1086–1118.

[8] J. Smoller, *Shock Wave and Reaction-Diffusion Equation*, Springer-Verlag, New York, 1983.

[9] L. A. Ying and Z. H. Teng, *Riemann problems for a reacting and convection hyperbolic system*, J. Approx. Theory, 1 (1984) pp. 95–122.

CLASSIFICATION OF THE RIEMANN PROBLEM FOR TWO-DIMENSIONAL GAS DYNAMICS*

CARSTEN W. SCHULZ-RINNE†

Abstract. The Riemann problem for two-dimensional gas dynamics with isentropic and polytropic gas is considered. The initial data is constant in each quadrant and chosen so that only a rarefaction wave, shock wave or slip line connects two neighboring constant initial states. With this restriction, the existence of sixteen (respectively, fifteen) genuinely different wave combinations for isentropic (respectively, polytropic) gas is proved. For each configuration the relations for the initial data and the symmetry properties of the solution are given. This paper corrects the conjectured classification presented in T. Zhang and Y. Zheng [*SIAM J. Math. Anal.*, 21 (1990), pp. 593–630].

Key words. Riemann problem, gas dynamics, initial data, compatibility conditions, self-similar solution

AMS(MOS) subject classifications. 35L65, 35L67, 76N15

1. Introduction. The study of the Riemann problem for gas dynamics has a long tradition, starting with the work of Riemann himself in the last century. In the last twenty years the Riemann problem for one-dimensional gas dynamics has been studied, and the results have been published in [6], [7], and [9] (they also contain further references). More recently the research was extended toward two-dimensional scalar conservation laws [2]–[4], [8], [10]. Riemann problems for two-dimensional gas dynamics were considered in [1] and [11].

Under certain assumptions, Zhang and Zheng [11] conjectured the existence of seventeen reasonable combinations of initial data (counting two subcases individually). Six of their configurations contain no slip lines. In this paper we analyze the same problem more thoroughly. We are able to prove that for isentropic gas one of these six configurations does not exist and one is centrally symmetric. For polytropic gas both cannot exist. Moreover, one of the remaining four configurations is always axially symmetric.

After exposing the problem in the following section, we classify the Riemann problem according to the combination of the elementary waves in §3. There it is shown that only sixteen (respectively, fifteen) genuinely different configurations for isentropic (respectively, polytropic) gas exist, compared to seventeen found by Zhang and Zheng [11]. These numbers are based on the same method of counting. In particular, those combinations which can be obtained by coordinate transformations are not counted.

Numerical solutions for each configuration have recently been computed by the author in joint work with Collins and Glaz. The wave structures are analyzed and illustrated by contour plots in [5].

2. Problem definition. The Euler equations of inviscid compressible isentropic flow consist of the continuity equations for the conservation of mass and momentum. For polytropic gas we have an additional equation for the conservation of energy.

*Received by the editors May 13, 1991; accepted for publication (in revised form) March 11, 1992.

†Seminar für Angewandte Mathematik, Eidgenössische Technische Hochschule Zürich, 8092 Zürich, Switzerland.

The conservation form of these equations in Cartesian coordinates, together with the equation of state, is

$$U_t + F(U)_x + G(U)_y = 0, \tag{2.1}$$

where

$$U = \begin{pmatrix} \rho \\ \rho u \\ \rho v \end{pmatrix}, \quad F(U) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \end{pmatrix}, \quad G(U) = \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \end{pmatrix}, \quad p = A\rho^\gamma$$

for isentropic gas and

$$U = \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ \rho E \end{pmatrix}, \quad F(U) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ u(\rho E + p) \end{pmatrix}, \quad G(U) = \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ v(\rho E + p) \end{pmatrix},$$

$$E = \frac{1}{(\gamma - 1)} \frac{p}{\rho} + \frac{u^2 + v^2}{2}$$

for polytropic gas. Here ρ is the density, u the x-velocity component, v the y-velocity component, p the pressure, E the energy, $\gamma > 1$ the ratio of specific heats of the gas, and $A > 0$ a constant.

The characteristic speeds of (2.1) in x- (or y-) direction, i.e., the eigenvalues of the Jacobian matrix $\nabla_U F$ (or $\nabla_U G$) are $\lambda_- = u - c$, $\lambda_0 = u$, and $\lambda_+ = u + c$ (or $\lambda_- = v - c$, $\lambda_0 = v$, and $\lambda_+ = v + c$). Here the sound speed c is defined by $c^2 = \gamma p/\rho$.

The Riemann problem in the (x, y)-plane is the initial value problem for (2.1) with initial data

$$(\rho, u, v)(x, y, 0) = (\rho_i, u_i, v_i), \qquad i = 1, \dots, 4 \tag{2.2}$$

for isentropic gas and

$$(p, \rho, u, v)(x, y, 0) = (p_i, \rho_i, u_i, v_i), \qquad i = 1, \dots, 4 \tag{2.2'}$$

for polytropic gas, where i denotes the ith quadrant.

The solution is a function of the similarity variables $\xi = x/t$ and $\eta = y/t$ and is called pseudostationary flow. Far enough away from the origin, the general solution consists of four planar waves, each parallel to one of the coordinate axes, between the four constant initial states. In general, a planar wave is formed by up to three elementary waves corresponding to the eigenvalues λ_-, λ_0, and λ_+: a backward rarefaction wave $\overleftarrow{R}$ or shock wave $\overleftarrow{S}$, a slip line (respectively, a contact discontinuity) for isentropic (respectively, polytropic) gas J, and a forward rarefaction wave $\overrightarrow{R}$ or shock wave $\overrightarrow{S}$. This study of the two-dimensional Riemann problem is restricted to situations where each planar wave consists of a single elementary wave. Thus the initial data has to be chosen so that only a rarefaction wave, shock wave or slip line[1] connects two neighboring constant initial states.

[1] From here on the use of the term *slip line* always denotes a slip line for isentropic gas and a contact discontinuity for polytropic gas.

Before we start with the classification of the two-dimensional Riemann problem, the formulas for the one-dimensional elementary waves between two constant states are briefly reviewed.

Across a backward (respectively, forward) rarefaction wave $\overleftarrow{R}$ (respectively, $\overrightarrow{R}$) the corresponding Riemann invariants are constant. They are $w + (2/(\gamma - 1))c$ (respectively, $w - (2/(\gamma - 1))c$) for isentropic gas and, additionally, the entropy s for polytropic gas, where w is the velocity. For a given left and right state (denoted by the indices l and r), we thus have

$$w_r - w_l = \frac{2\sqrt{\gamma}}{\gamma - 1}\left(\sqrt{\frac{p_l}{\rho_l}} - \sqrt{\frac{p_r}{\rho_r}}\right) =: \Phi_{lr} \qquad \text{(respectively, } w_l - w_r = \Phi_{lr}\text{)}$$

for $\overleftarrow{R}$ (respectively, $\overrightarrow{R}$). For polytropic gas we further find

$$\frac{p_l}{p_r} = \left(\frac{\rho_l}{\rho_r}\right)^{\gamma},$$

using the additional Riemann invariant s.

The Rankine–Hugoniot conditions for the system of equations (2.1) give the relations between the states on each side of a shock wave S. From these conditions we can derive

$$(w_l - w_r)^2 = \frac{(p_l - p_r)(\rho_l - \rho_r)}{\rho_l \rho_r} =: \Psi_{lr}^2 \qquad (\Psi_{lr} > 0)$$

and, additionally,

$$\frac{\rho_l}{\rho_r} = \left(\frac{p_l}{p_r} + \frac{(\gamma - 1)}{(\gamma + 1)}\right) \Bigg/ \left(1 + \frac{(\gamma - 1)}{(\gamma + 1)}\frac{p_l}{p_r}\right) =: \Pi_{lr}$$

for polytropic gas.

The type of elementary wave is determined by the pressure and the velocity inequalities:

$$\begin{array}{llcllc} p_l < p_r, & w_l < w_r : & \overrightarrow{R}, & p_l > p_r, & w_l < w_r : & \overleftarrow{R}, \\ p_l < p_r, & w_l > w_r : & \overleftarrow{S}, & p_l > p_r, & w_l > w_r : & \overrightarrow{S}. \end{array}$$

Across a contact discontinuity J the (normal) velocity and the pressure are constant, but the density can jump arbitrarily. For isentropic gas a slip line J only occurs in two-dimensional flow when the density and the normal velocity are constant and the tangential velocity is discontinuous.

3. Classification. In the following we assume that the initial data (2.2) or (2.2') are chosen so that only one elementary wave connects two neighboring constant states. First we consider all combinations that involve rarefaction and shock waves exclusively. Thereafter, all combinations involving slip lines are considered. For all possible configurations we give the relations that have to be satisfied by the initial data and the symmetry properties of the solution.

Configurations without slip lines. Here we consider the configurations that only involve rarefaction and shock waves. Then there exist only three distinct relations between the pressure values in the four quadrants:

$$\begin{array}{ccc} p_2 & < & p_1 \\ \vee & & \vee \\ p_3 & < & p_4 \end{array} \qquad \begin{array}{ccc} p_2 & < & p_1 \\ \wedge & & \vee \\ p_3 & > & p_4 \end{array} \qquad \begin{array}{ccc} p_2 & < & p_1 \\ \vee & & \vee \\ p_3 & > & p_4 \end{array}$$

The remaining relations can be derived from the above by coordinate transformations.

For the velocities four different relations are possible:

$$\begin{array}{cccc}
u_2 = u_3 < u_4 = u_1 & u_2 = u_3 > u_4 = u_1 & u_2 = u_3 < u_4 = u_1 & u_2 = u_3 > u_4 = u_1 \\
\text{and} & \text{and} & \text{and} & \text{and} \\
v_1 = v_2 > v_3 = v_4 & v_1 = v_2 < v_3 = v_4 & v_1 = v_2 < v_3 = v_4 & v_1 = v_2 > v_3 = v_4.
\end{array}$$

Altogether we get twelve configurations:

$p_1 > p_2,\ p_4 > p_3$:	$\overrightarrow{R}_{21}\overrightarrow{R}_{32}\overrightarrow{R}_{34}\overrightarrow{R}_{41}$	$\overleftarrow{S}_{21}\overleftarrow{S}_{32}\overleftarrow{S}_{34}\overleftarrow{S}_{41}$	$\overrightarrow{R}_{21}\overleftarrow{S}_{32}\overrightarrow{R}_{34}\overleftarrow{S}_{41}$	$\overleftarrow{S}_{21}\overrightarrow{R}_{32}\overleftarrow{S}_{34}\overrightarrow{R}_{41}$
$p_1 > p_2,\ p_4 < p_3$:	$\overrightarrow{R}_{21}\overleftarrow{R}_{32}\overleftarrow{R}_{34}\overrightarrow{R}_{41}$	$\overleftarrow{S}_{21}\overrightarrow{S}_{32}\overrightarrow{S}_{34}\overleftarrow{S}_{41}$	$\overrightarrow{R}_{21}\overrightarrow{S}_{32}\overleftarrow{R}_{34}\overleftarrow{S}_{41}$	$\overleftarrow{S}_{21}\overleftarrow{R}_{32}\overrightarrow{S}_{34}\overrightarrow{R}_{41}$
$p_1 > p_2 > p_3 > p_4$:	$\overrightarrow{R}_{21}\overrightarrow{R}_{32}\overleftarrow{R}_{34}\overrightarrow{R}_{41}$	$\overleftarrow{S}_{21}\overleftarrow{S}_{32}\overrightarrow{S}_{34}\overleftarrow{S}_{41}$	$\overrightarrow{R}_{21}\overleftarrow{S}_{32}\overleftarrow{R}_{34}\overleftarrow{S}_{41}$	$\overleftarrow{S}_{21}\overrightarrow{R}_{32}\overrightarrow{S}_{34}\overrightarrow{R}_{41}$

In this table and in the following E_{ij} with $E \in \{J, \overleftarrow{R}, \overrightarrow{R}, \overleftarrow{S}, \overrightarrow{S}\}$ and $i, j \in \{1, 2, 3, 4\}$ denotes an elementary wave E between the ith and jth quadrant.

Obviously, exchanging the axes in the right column gives the neighboring one. Examination of the configurations in the last row shows that they are impossible.

For $\overrightarrow{R}_{21}\overrightarrow{R}_{32}\overleftarrow{R}_{34}\overrightarrow{R}_{41}$ we have $v_4 - v_1 = \Phi_{41}$, $v_3 - v_2 = \Phi_{32}$, $v_2 = v_1$, $v_3 = v_4$. This implies $\sqrt{p_2/\rho_2} - \sqrt{p_1/\rho_1} = \sqrt{p_3/\rho_3} - \sqrt{p_4/\rho_4}$ in contradiction to the pressure inequality.

For $\overleftarrow{S}_{21}\overleftarrow{S}_{32}\overrightarrow{S}_{34}\overleftarrow{S}_{41}$ and $\overrightarrow{R}_{21}\overleftarrow{S}_{32}\overleftarrow{R}_{34}\overleftarrow{S}_{41}$ we have $v_4 - v_1 = \Psi_{41}$, $v_3 - v_2 = \Psi_{32}$, $v_2 = v_1$, $v_3 = v_4$ yielding $\Psi_{41} = \Psi_{32}$. Since the pressure inequality gives $p_4 - p_1 < p_3 - p_2$ and $1/\rho_1 - 1/\rho_4 < 1/\rho_2 - 1/\rho_3$, this is a contradiction, too.

At this point six configurations are remaining which are examined individually in the following.

Configuration 1. $\overrightarrow{R}_{21}\overrightarrow{R}_{32}\overrightarrow{R}_{34}\overrightarrow{R}_{41}$.

We have

$$p_1 > p_2,\ p_4 > p_3$$

and

$$\begin{array}{llll}
u_2 - u_1 = \Phi_{21}, & u_3 - u_4 = \Phi_{34}, & u_3 = u_2, & u_4 = u_1, \\
v_4 - v_1 = \Phi_{41}, & v_3 - v_2 = \Phi_{32}, & v_2 = v_1, & v_3 = v_4.
\end{array}$$

This gives the so-called compatibility condition $\Phi_{21} = \Phi_{34}$. For polytropic gas we have to include the following equations:

$$\frac{\rho_i}{\rho_j} = \left(\frac{p_i}{p_j}\right)^{1/\gamma} \quad \text{for } (i,j) \in \{(2,1),\ (3,4),\ (3,2),\ (4,1)\}.$$

Configuration 2. $\overrightarrow{R}_{21}\overleftarrow{R}_{32}\overleftarrow{R}_{34}\overrightarrow{R}_{41}$.

We have

$$p_1 > p_2,\ p_4 < p_3$$

and

$$\begin{array}{llll}
u_2 - u_1 = \Phi_{21}, & u_4 - u_3 = \Phi_{34}, & u_3 = u_2, & u_4 = u_1, \\
v_4 - v_1 = \Phi_{41}, & v_2 - v_3 = \Phi_{32}, & v_2 = v_1, & v_3 = v_4,
\end{array}$$

so that the compatibility conditions are $\Phi_{21} = -\Phi_{34}$ and $\Phi_{41} = -\Phi_{32}$. For polytropic gas we append the same equations as in Configuration 1.

Thus we must have $p_1 = p_3$ and $p_2 = p_4$ implying $u_1 - u_2 = v_1 - v_4$ and $u_4 - u_3 = v_2 - v_3$. Consequently, the solutions are symmetric to $\eta - \xi = v_1 - u_1$ and $\xi + \eta = u_2 + v_2$.

Configuration 3. $\overleftarrow{S}_{21}\overleftarrow{S}_{32}\overleftarrow{S}_{34}\overleftarrow{S}_{41}$.

We have

$$p_1 > p_2, p_4 > p_3 \tag{3.1}$$

and

$$\begin{aligned} u_2 - u_1 &= \Psi_{21}, & u_3 - u_4 &= \Psi_{34}, & u_3 &= u_2, & u_4 &= u_1, \\ v_4 - v_1 &= \Psi_{41}, & v_3 - v_2 &= \Psi_{32}, & v_2 &= v_1, & v_3 &= v_4. \end{aligned}$$

This gives the compatibility conditions

$$\Psi_{21} = \Psi_{34} \quad \text{and} \quad \Psi_{41} = \Psi_{32}. \tag{3.2}$$

For polytropic gas the following equations must be added:

$$\frac{\rho_i}{\rho_j} = \Pi_{ij} \quad \text{for } (i,j) \in \{(2,1),\ (3,4),\ (3,2),\ (4,1)\}. \tag{3.3}$$

Due to the compatibility conditions we have to choose $p_4 = p_2$ (which implies $\rho_4 = \rho_2$) according to the Theorem 1. Then the compatibility conditions (3.2) become a single equation, and we have $u_2 - u_1 = v_4 - v_1$. Consequently, the solutions are symmetric to $\eta - \xi = v_1 - u_1$.

THEOREM 1. *The inequality* (3.1), *the compatibility conditions* (3.2) *and the additional equations for polytropic gas* (3.3) *can only be satisfied if* $p_4 = p_2$. (*For polytropic gas it is assumed that* $1 < \gamma \leq 3$ *holds.*)

Proof. For isentropic gas we apply the equation of state to the compatibility conditions (3.2) and set

$$\rho_1 = x\rho_3, \quad \rho_2 = y\rho_3, \quad \text{and} \quad \rho_4 = z\rho_3,$$

getting

$$(x^\gamma - y^\gamma)\left(\frac{1}{y} - \frac{1}{x}\right) = (z^\gamma - 1)\left(1 - \frac{1}{z}\right), \qquad (x^\gamma - z^\gamma)\left(\frac{1}{z} - \frac{1}{x}\right) = (y^\gamma - 1)\left(1 - \frac{1}{y}\right).$$

We define $f(z, y)$ as the difference of the left- and right-hand side of the last equation (assuming x to be fixed). Now we have to prove that $f(y, z) = 0$ and $f(z, y) = 0$ for $x > y$, $z > 1$ only if $y = z$.

We find that

$$z\left(1 - \frac{z}{x}\right) f(y, z) + z(1 - z) f(z, y)$$

is a quadratic polynomial in z with the roots

$$z_1(y) = \frac{xy}{y(x+1) - x} \quad \text{and} \quad z_2(y) = \frac{x(x^\gamma - 1)}{y^\gamma(x-1) + x^\gamma - x}.$$

By construction, these roots are the only candidates for the roots of $f(y,z)$ and $f(z,y)$. After replacing z by z_1 in $f(y,z)$ and $f(z,y)$, we multiply both of them by a positive factor that has no influence on the roots of a function:

$$\begin{aligned} g_1(y) &= \frac{(y(x+1)-x)^\gamma}{y^\gamma(y^{-1}-x^{-1})} f(y,z_1) = \frac{(y(x+1)-x)^\gamma}{y^\gamma(1-y^{-1})} \cdot f(z_1,y) \\ &= \left(\frac{1+x^\gamma}{y^\gamma}-1\right)(y(x+1)-x)^\gamma - x^\gamma. \end{aligned}$$

Since $g_1(1) = g_1(x) = 0$ and $g_1(y)$ has its extremum for $y^{\gamma+1} = x(x^\gamma+1)/(x+1)$, there exists no root $y \in (1,x)$ of $g_1(y)$ and, consequently, of $f(y,z)$ or $f(z,y)$.

Now we define

$$\begin{aligned} g_2(y) &= (y^\gamma(x-1)+x^\gamma-x)^\gamma\,[f(y,z_2)+f(z_2,y)] \\ &= (y^\gamma(x-1)+x^\gamma-x)^\gamma\left(\frac{1}{y}(x^\gamma-1)-x^{\gamma-1}+1\right) - x^\gamma(x^\gamma-1)^\gamma\left(1-\frac{1}{x}\right). \end{aligned}$$

The common roots of $f(y,z_2)$ and $f(z_2,y)$ are a subset of the roots of $g_2(y)$. As before, we have $g_2(1) = g_2(x) = 0$, and we compute the first derivative of $g_2(y)$:

$$\begin{aligned} g_2'(y) =& \gamma^2 y^{\gamma-1}(x-1)(y^\gamma(x-1)+x^\gamma-x)^{\gamma-1}\left(\frac{1}{y}(x^\gamma-1)-x^{\gamma-1}+1\right) \\ &- (y^\gamma(x-1)+x^\gamma-x)^\gamma\frac{1}{y^2}(x^\gamma-1). \end{aligned}$$

In order to show that $g_2'(y)$ has at most two roots in $(1,x)$, we examine

$$\widehat{g_2}(y) = y^2(y^\gamma(x-1)+x^\gamma-x)^{1-\gamma}g_2'(y),$$

which has the same roots as $g_2'(y)$ in $(1,x)$. $\widehat{g_2}(y)$ is only extremal at $y = (1-\gamma^{-1})(x^\gamma - 1)/(x^{\gamma-1}-1)$. Thus $g_2'(y)$ has no more than two roots, and, consequently, $g_2(y)$ has at most one root in $(1,x)$.

On the other hand, $f(1,1)\,f(x,x) < 0$ and $f(y,y)$ is smooth, so that a root of $f(y,y)$ does exist in $(1,x)$. This must be the unique root of $g_2(y)$ in $(1,x)$, showing that $y = z$ is required as claimed. Moreover, for given x we can compute y by solving $y = z_2(y)$.

For polytropic gas we rewrite the compatibility conditions (3.2) to get

$$\begin{aligned} \frac{\rho_3}{\rho_2}(p_1-p_2)\left(1-\frac{\rho_2}{\rho_1}\right) &= (p_4-p_3)\left(1-\frac{\rho_3}{\rho_4}\right), \\ \frac{\rho_3}{\rho_4}(p_1-p_4)\left(1-\frac{\rho_4}{\rho_1}\right) &= (p_2-p_3)\left(1-\frac{\rho_3}{\rho_2}\right). \end{aligned}$$

Then we use the additional equations (3.3) to eliminate ρ_i, $i = 1,\dots,4$ in these equations. Setting

$$p_1 = xp_3, \quad p_2 = yp_3, \quad p_4 = zp_3 \quad \text{and} \quad \varepsilon = \frac{(\gamma-1)}{(\gamma+1)},$$

the compatibility conditions are equivalent to

$$(x-y)^2(1+\varepsilon y)(z+\varepsilon) = (z-1)^2(y+\varepsilon)(x+\varepsilon y),$$
$$(x-z)^2(1+\varepsilon z)(y+\varepsilon) = (y-1)^2(z+\varepsilon)(x+\varepsilon z).$$

We define $f(z,y)$ as the difference of the left- and right-hand side of the last equation (assuming x to be fixed). Now we have to prove that $f(y,z)=0$ and $f(z,y)=0$ for $x>y$, $z>1$ only if $y=z$.

The resultant $r(z)$ of $f(y,z)$ and $f(z,y)$ has the following roots:

$$z_0 = -\varepsilon, \quad z_1 = x, \quad z_2 = 1,$$
$$z_3^\pm = \frac{1}{2(1+2\varepsilon)}\left(\varepsilon(x+1) \pm \sqrt{\varepsilon^2(x+1)^2 + 4(1+2\varepsilon)x}\right),$$
$$z_4^\pm = \frac{1}{2\varepsilon(2\varepsilon^2+\varepsilon x - x)}\Bigl(-2\varepsilon^3 + (1+\varepsilon-6\varepsilon^2+2\varepsilon^3)x + (1-3\varepsilon+2\varepsilon^2)x^2$$
$$\pm\sqrt{x-1}\sqrt{2\varepsilon^3(x-1)+(1-\varepsilon-2\varepsilon^2)x}$$
$$\cdot\sqrt{-(1+3\varepsilon-2\varepsilon^2-2\varepsilon^3x^{-1})x-(1-\varepsilon)x^2}\Bigr).$$

By the definition of the resultant, the common roots of $f(y,z)$ and $f(z,y)$ are a subset of the roots of $r(z)$. Obviously, z_0, z_1, z_2 and z_3^- are not in the interval $(1,x)$. Under the assumption for γ, we have $0 < \varepsilon \leq \frac{1}{2}$. Then $z_4^\pm$ is well defined and its discriminant is negative because $x-1$ and $2\varepsilon^3(x-1)+(1-\varepsilon-2\varepsilon^2)x$ are positive and $-(1+3\varepsilon-2\varepsilon^2-2\varepsilon^3x^{-1})x-(1-\varepsilon)x^2$ is negative since

$$1+3\varepsilon-2\varepsilon^2-2\varepsilon^3x^{-1} \geq 1+3\varepsilon-2\varepsilon-2\varepsilon = 1-\varepsilon > 0.$$

Thus $z_4^\pm$ is complex, and z_3^+ is the only positive root of $r(z)$. It is easy to verify that $z_3^+ \in (1,x)$ and that $f(z_3^+, z_3^+) = 0$. □

Configuration 4. $\overleftarrow{S}_{21}\overrightarrow{S}_{32}\overrightarrow{S}_{34}\overleftarrow{S}_{41}$.

We have

$$p_1 > p_2, p_4 < p_3$$

and the same equations and compatibility conditions as in Configuration 3.

Necessarily we must have $p_1 = p_3$ and $p_2 = p_4$ (which implies $\rho_1 = \rho_3$ and $\rho_2 = \rho_4$) yielding $u_2 - u_1 = v_4 - v_1$ and $u_3 - u_4 = v_3 - v_2$. Consequently, the solutions are symmetric to $\eta - \xi = v_1 - u_1$ and $\xi + \eta = u_2 + v_2$.

Configuration 5. $\overrightarrow{R}_{21}\overleftarrow{S}_{32}\overrightarrow{R}_{34}\overleftarrow{S}_{41}$.

We have

$$p_1 > p_2,\ p_4 > p_3 \tag{3.4}$$

and

$$u_2 - u_1 = \Phi_{21}, \quad u_3 - u_4 = \Phi_{34}, \quad u_3 = u_2, \quad u_4 = u_1,$$
$$v_4 - v_1 = \Psi_{41}, \quad v_3 - v_2 = \Psi_{32}, \quad v_2 = v_1, \quad v_3 = v_4.$$

This gives the compatibility conditions

$$\Phi_{21} = \Phi_{34} \quad \text{and} \quad \Psi_{41} = \Psi_{32}. \tag{3.5}$$

For polytropic gas the equations

$$\frac{\rho_3}{\rho_2} = \Pi_{32}, \quad \frac{\rho_4}{\rho_1} = \Pi_{41}, \quad \frac{\rho_2}{\rho_1} = \left(\frac{p_2}{p_1}\right)^{1/\gamma}, \quad \text{and} \quad \frac{\rho_3}{\rho_4} = \left(\frac{p_3}{p_4}\right)^{1/\gamma} \tag{3.6}$$

are added.

We can prove that this configuration is impossible.

THEOREM 2. *There exist no p_i, $i = 1,\dots,4$ satisfying the inequality* (3.4), *the compatibility conditions* (3.5) *and the additional equations for polytropic gas* (3.6).

Proof. For isentropic gas we apply the equation of state to the compatibility conditions (3.5), getting

$$\sqrt{\rho_1^{\gamma-1}} - \sqrt{\rho_2^{\gamma-1}} = \sqrt{\rho_4^{\gamma-1}} - \sqrt{\rho_3^{\gamma-1}}, \qquad \frac{(\rho_1^\gamma - \rho_4^\gamma)(\rho_1 - \rho_4)}{\rho_1\rho_4} = \frac{(\rho_2^\gamma - \rho_3^\gamma)(\rho_2 - \rho_3)}{\rho_2\rho_3}.$$

Now we show that for any p_i, $i = 1,\dots,4$ satisfying the inequality (3.4) and the first compatibility condition, the second is violated. Defining

$$\delta = \frac{2}{(\gamma-1)} \quad \text{and} \quad R_i = \rho_i^{1/\delta}, \quad i = 1,\dots,4,$$

the first compatibility condition is equivalent to $R_1 - R_2 = R_4 - R_3$. Introducing $\Delta = R_1 - R_2 > 0$, we have

$$R_1 = R_2 + \Delta \quad \text{and} \quad R_4 = R_3 + \Delta. \tag{3.7}$$

Then the second compatibility condition can be written as

$$\frac{[(R_2+\Delta)^{\gamma\delta} - (R_3+\Delta)^{\gamma\delta}][(R_2+\Delta)^\delta - (R_3+\Delta)^\delta]}{(R_2+\Delta)^\delta (R_3+\Delta)^\delta} = \frac{[R_2^{\gamma\delta} - R_3^{\gamma\delta}][R_2^\delta - R_3^\delta]}{R_2^\delta R_3^\delta}.$$

We define $f(\Delta)$ as the difference of the left- and right-hand side of the last equation. Obviously, $f(0) = 0$. Now we have to prove that $f(\Delta)$ does not vanish for any positive R_2, R_3 and Δ. We differentiate f with respect to Δ. Using $\gamma\delta = \delta + 2$ and (3.7), this yields

$$\begin{aligned}\frac{\partial f}{\partial \Delta} = &\frac{1}{R_1^{\delta+1}R_4^{\delta+1}} \{(\delta+2)R_1R_4[R_1^{\delta+1} - R_4^{\delta+1}][R_1^\delta - R_4^\delta] \\ &- \delta[R_1^{\delta+2} - R_4^{\delta+2}][R_1^{\delta+1} - R_4^{\delta+1}]\}.\end{aligned}$$

Introducing $R = R_1/R_4 = (R_2 + \Delta)/(R_3 + \Delta) > 1$, this derivative becomes

$$\frac{\partial f}{\partial \Delta} = \frac{R_4}{R^{\delta+1}} (R^{\delta+1} - 1) \underbrace{\{(\delta+2)R(R^\delta - 1) - \delta(R^{\delta+2} - 1)\}}_{=h(R)}.$$

In order to show that $\partial f/\partial\Delta$ is negative for all positive R_2, R_3, and Δ, it is equivalent to show that the function $h(R)$ defined above is negative for all R strictly greater than one. Since $h(1) = 0$ we differentiate h with respect to R and get

$$h'(R) = (\delta+2)\{(\delta+1)R^\delta - 1 - \delta R^{\delta+1}\}.$$

Observing that $h'(1) = 0$, we compute $h''(R)$ to find

$$h''(R) = (\delta+2)(\delta+1)\delta R^{\delta-1}\{1 - R\}.$$

Obviously, $h''(R)$ is negative for R strictly greater than one. Thus we have proved that $f(\Delta)$ is negative for all positive R_2, R_3 and Δ.

For polytropic gas we use the additional equations (3.6) to derive the following equation, which is independent of ρ_i, $i = 1, \dots, 4$:

$$\left(\frac{p_2}{p_3}\right)^{1/\gamma} \Pi_{32} = \left(\frac{p_1}{p_4}\right)^{1/\gamma} \Pi_{41}.$$

Introducing

$$f(P) = \left(\frac{1}{P}\right)^{1/\gamma} \left(P + \frac{(\gamma-1)}{(\gamma+1)}\right) \Big/ \left(1 + \frac{(\gamma-1)}{(\gamma+1)} P\right),$$

this can be written as

$$f\left(\frac{p_3}{p_2}\right) = f\left(\frac{p_4}{p_1}\right).$$

Since $f(P)$ is strictly decreasing for $P \in (0, \infty)$, the last equation implies that

$$\text{(3.8)} \qquad \frac{p_3}{p_2} = \frac{p_4}{p_1}, \quad \text{which is equivalent to} \quad \frac{p_3}{p_4} = \frac{p_2}{p_1}.$$

Again using the additional equations, we eliminate ρ_i, $i = 1, \dots, 4$ in the first compatibility condition and get

$$\sqrt{\Pi_{41}} \left(1 - \left(\frac{p_2}{p_1}\right)^{(\gamma-1)/2\gamma}\right) = \sqrt{\frac{p_4}{p_1}} \left(1 - \left(\frac{p_3}{p_4}\right)^{(\gamma-1)/2\gamma}\right).$$

With our previous result (3.8), it follows that this is equivalent to $\Pi_{41} = p_4/p_1$. The only admissible solution $p_1 = p_4$ violates the inequality (3.4). □

Configuration 6. $\overrightarrow{R}_{21}\overrightarrow{S}_{32}\overleftarrow{R}_{34}\overleftarrow{S}_{41}$.

We have

$$\text{(3.9)} \qquad p_1 > p_2, \, p_4 < p_3$$

and

$$\begin{aligned} u_2 - u_1 &= \Phi_{21}, & u_4 - u_3 &= \Phi_{34}, & u_3 &= u_2, & u_4 &= u_1, \\ v_4 - v_1 &= \Psi_{41}, & v_3 - v_2 &= \Psi_{32}, & v_2 &= v_1, & v_3 &= v_4, \end{aligned}$$

so that the compatibility conditions are

$$\text{(3.10)} \qquad \Phi_{21} = -\Phi_{34} \quad \text{and} \quad \Psi_{41} = \Psi_{32}.$$

For polytropic gas we have the same additional equations as in Configuration 5, and this configuration is impossible, too.

THEOREM 3. *For polytropic gas there exist no p_i, $i = 1, \dots, 4$ satisfying the inequality* (3.9), *the compatibility conditions* (3.10), *and the additional equations for polytropic gas* (3.6).

Proof. As in the preceding proof, we use the additional equations (3.6) to derive that $p_3/p_2 = p_4/p_1$. This is in contradiction to the inequality (3.9), which implies that $p_3/p_2 > 1 > p_4/p_1$. □

For isentropic gas the following theorem states that we have to choose $p_3 = p_1$ and $p_4 = p_2$ (which implies $\rho_3 = \rho_1$ and $\rho_4 = \rho_2$). Thus the solutions are symmetric with respect to the point $(\xi, \eta) = \left(\frac{1}{2}(u_1 + u_2), \frac{1}{2}(v_1 + v_3)\right)$.

THEOREM 4. *For isentropic gas the inequality* (3.9) *and the compatibility conditions* (3.10) *can only be satisfied if* $p_3 = p_1$ *and* $p_4 = p_2$.

Proof. We apply the equation of state to the compatibility conditions (3.10), getting

$$\sqrt{\rho_1^{\gamma-1}} - \sqrt{\rho_2^{\gamma-1}} = \sqrt{\rho_3^{\gamma-1}} - \sqrt{\rho_4^{\gamma-1}}, \qquad \frac{(\rho_1^\gamma - \rho_4^\gamma)(\rho_1 - \rho_4)}{\rho_1\rho_4} = \frac{(\rho_3^\gamma - \rho_2^\gamma)(\rho_3 - \rho_2)}{\rho_3\rho_2}.$$

Defining

$$\delta = \frac{2}{(\gamma - 1)} \quad \text{and} \quad R_i = \rho_i^{1/\delta}, \quad i = 1, \ldots, 4,$$

the first compatibility condition becomes $R_1 - R_2 = R_3 - R_4$, which is equivalent to $R_1 - R_3 = R_2 - R_4$. Introducing $\Delta = R_1 - R_3$, we have $R_1 = R_3 + \Delta$ and $R_4 = R_2 - \Delta$. Then the second compatibility condition can be written as

$$\frac{[(R_3 + \Delta)^{\gamma\delta} - (R_2 - \Delta)^{\gamma\delta}][(R_3 + \Delta)^\delta - (R_2 - \Delta)^\delta]}{(R_3 + \Delta)^\delta (R_2 - \Delta)^\delta} = \frac{[R_3^{\gamma\delta} - R_2^{\gamma\delta}][R_3^\delta - R_2^\delta]}{R_3^\delta R_2^\delta}.$$

We define $f(\Delta)$ as the difference of the left- and right-hand side of the last equation. Obviously, $f(0) = 0$. Now we prove that $f(\Delta)$ is strictly increasing for any positive R_2, R_3 and $\Delta \in (-R_3, R_2)$. We differentiate f with respect to Δ. Using $\gamma\delta = \delta + 2$ and introducing $R = (R_3 + \Delta)/(R_2 - \Delta) = R_1/R_4 > 1$, this yields

$$\frac{\partial f}{\partial \Delta} = \frac{R_4}{R^{\delta+1}} (R^{\delta+1} + 1) \{(\delta + 2)R(R^\delta - 1) + \delta (R^{\delta+2} - 1)\} > 0.$$

Hence f is strictly increasing and $\Delta = 0$ is its unique root, i.e., $p_3 = p_1$ and $p_4 = p_2$ is necessary to satisfy the compatibility conditions. □

Configurations involving slip lines. Now we consider all combinations involving slip lines J. There are two genuinely different configurations with four J's, one where all the J's are moving clockwise and one where two J's are moving in the opposite direction.

Three J's imply $p_1 = p_2 = p_3 = p_4$ in contradiction to the pressure inequality of the fourth wave.

Two J's are either neighbors or not. In the first case we assume that the J's are between the third quadrant and its neighbors. Then we find the following eight configurations:

$$p_1 > p_2 = p_3 = p_4\colon \quad \overrightarrow{R}_{21}J_{32}J_{34}\overrightarrow{R}_{41} \quad \overleftarrow{S}_{21}J_{32}J_{34}\overleftarrow{S}_{41} \quad \overrightarrow{R}_{21}J_{32}J_{34}\overleftarrow{S}_{41} \quad \overleftarrow{S}_{21}J_{32}J_{34}\overrightarrow{R}_{41}$$

$$p_1 < p_2 = p_3 = p_4\colon \quad \overleftarrow{R}_{21}J_{32}J_{34}\overleftarrow{R}_{41} \quad \overrightarrow{S}_{21}J_{32}J_{34}\overrightarrow{S}_{41} \quad \overleftarrow{R}_{21}J_{32}J_{34}\overrightarrow{S}_{41} \quad \overrightarrow{S}_{21}J_{32}J_{34}\overleftarrow{R}_{41}.$$

As before, exchanging the axes in the right column gives the neighboring one.

In the case where the J's are not neighbors, three different combinations are possible:

$$p_1 = p_2 > p_3 = p_4\colon \quad J_{21}\overrightarrow{R}_{32}J_{34}\overrightarrow{R}_{41} \quad J_{21}\overleftarrow{S}_{32}J_{34}\overleftarrow{S}_{41} \quad J_{21}\overleftarrow{S}_{32}J_{34}\overrightarrow{R}_{41}.$$

A case with one J would imply $u_1 = u_2 = u_3 = u_4$ or $v_1 = v_2 = v_3 = v_4$ in contradiction to the existence of three shock and rarefaction waves.

In the following, these eleven configurations involving slip lines are listed with their relations for the initial data and their symmetry properties. For polytropic gas we have to include additional equations. Namely, for a rarefaction or a shock wave between the ith and jth quadrant (i, $j \in \{1, \ldots, 4\}$) we add

$$\frac{\rho_i}{\rho_j} = \left(\frac{p_i}{p_j}\right)^{1/\gamma} \quad \text{or} \quad \frac{\rho_i}{\rho_j} = \Pi_{ij},$$

respectively.

Configuration A. $\overleftrightarrow{J_{21}J_{32}J_{34}J_{41}}$ (motion in opposite directions).
We have $p_1 = p_2 = p_3 = p_4$ and

$$u_1 = u_2 < u_3 = u_4, \qquad v_1 = v_4 < v_3 = v_2.$$

The solutions are symmetric with respect to the point $(\xi, \eta) = \left(\frac{1}{2}(u_1 + u_3), \frac{1}{2}(v_1 + v_2)\right)$ for isentropic gas.

Configuration B. $\overrightarrow{J_{21}J_{32}J_{34}J_{41}}$ (clockwise motion).
We have $p_1 = p_2 = p_3 = p_4$ and

$$u_1 = u_2 > u_3 = u_4, \qquad v_1 = v_4 < v_3 = v_2.$$

The solutions have the same symmetry properties as in Configuration A.

Configuration C. $\overrightarrow{R}_{21}J_{32}J_{34}\overrightarrow{R}_{41}$.
We have $p_1 > p_2 = p_3 = p_4$ and

$$u_2 - u_1 = \Phi_{21}, \quad u_3 = u_4 = u_1, \quad v_4 - v_1 = \Phi_{41}, \quad v_3 = v_2 = v_1.$$

The solutions are symmetric to $\eta - \xi = v_1 - u_1$.

Configuration D. $\overleftarrow{R}_{21}J_{32}J_{34}\overleftarrow{R}_{41}$.
We have $p_1 < p_2 = p_3 = p_4$ and

$$u_1 - u_2 = \Phi_{21}, \quad u_3 = u_4 = u_1, \quad v_1 - v_4 = \Phi_{41}, \quad v_3 = v_2 = v_1.$$

The solutions are symmetric to $\eta - \xi = v_1 - u_1$.

Configuration E. $\overleftarrow{S}_{21}J_{32}J_{34}\overleftarrow{S}_{41}$.
We have $p_1 > p_2 = p_3 = p_4$ and

$$u_2 - u_1 = \Psi_{21}, \quad u_3 = u_4 = u_1, \quad v_4 - v_1 = \Psi_{41}, \quad v_3 = v_2 = v_1.$$

The solutions are symmetric to $\eta - \xi = v_1 - u_1$.

Configuration F. $\overrightarrow{S}_{21}J_{32}J_{34}\overrightarrow{S}_{41}$.
We have $p_1 < p_2 = p_3 = p_4$ and the same equations as in Configuration E.

Configuration G. $\overrightarrow{R}_{21}J_{32}J_{34}\overleftarrow{S}_{41}$.
We have $p_1 > p_2 = p_3 = p_4$ and

$$u_2 - u_1 = \Phi_{21}, \quad u_3 = u_4 = u_1, \quad v_4 - v_1 = \Psi_{41}, \quad v_3 = v_2 = v_1.$$

Configuration H. $\overleftarrow{R}_{21}J_{32}J_{34}\overrightarrow{S}_{41}$.
We have $p_1 < p_2 = p_3 = p_4$ and

$$u_1 - u_2 = \Phi_{21}, \quad u_3 = u_4 = u_1, \quad v_4 - v_1 = \Psi_{41}, \quad v_3 = v_2 = v_1.$$

Configuration I. $J_{21}\overrightarrow{R}_{32}J_{34}\overrightarrow{R}_{41}$.
We have $p_1 = p_2 > p_3 = p_4$ and

$$u_1 = u_2 = u_3 = u_4, \quad v_4 - v_1 = \Phi_{41}, \quad v_3 - v_2 = \Phi_{32}.$$

Configuration J. $J_{21}\overleftarrow{S}_{32}J_{34}\overleftarrow{S}_{41}$.
We have $p_1 = p_2 > p_3 = p_4$ and

$$u_1 = u_2 = u_3 = u_4, \quad v_4 - v_1 = \Psi_{41}, \quad v_3 - v_2 = \Psi_{32}.$$

Configuration K. $J_{21}\overleftarrow{S}_{32}J_{34}\overrightarrow{R}_{41}$.
We have $p_1 = p_2 > p_3 = p_4$ and

$$u_1 = u_2 = u_3 = u_4, \quad v_4 - v_1 = \Phi_{41}, \quad v_3 - v_2 = \Psi_{32}.$$

Conclusion. Combining the results of this section, we have sixteen (respectively, fifteen) different configurations for the Riemann problem for isentropic (respectively, polytropic) gas in two space dimensions:

$$\begin{array}{llll}
4R: & \overrightarrow{R}_{21}\overrightarrow{R}_{32}\overrightarrow{R}_{34}\overrightarrow{R}_{41} & \overrightarrow{R}_{21}\overleftarrow{R}_{32}\overleftarrow{R}_{34}\overrightarrow{R}_{41} & \\
4S: & \overleftarrow{S}_{21}\overleftarrow{S}_{32}\overleftarrow{S}_{34}\overleftarrow{S}_{41} & \overleftarrow{S}_{21}\overrightarrow{S}_{32}\overrightarrow{S}_{34}\overleftarrow{S}_{41} & \\
2R+2S: & & \overrightarrow{R}_{21}\overrightarrow{S}_{32}\overleftarrow{R}_{34}\overleftarrow{S}_{41} & \text{(only for isentropic gas)} \\
4J: & \overleftrightarrow{J_{21}\,J_{32}\,J_{34}\,J_{41}} & \overrightarrow{J_{21}\,J_{32}\,J_{34}\,J_{41}} & \\
2J+2R: & \overrightarrow{R}_{21}\,J_{32}\,J_{34}\overrightarrow{R}_{41} & \overleftarrow{R}_{21}\,J_{32}\,J_{34}\overleftarrow{R}_{41} & J_{21}\overrightarrow{R}_{32}\,J_{34}\overrightarrow{R}_{41} \\
2J+2S: & \overleftarrow{S}_{21}\,J_{32}\,J_{34}\overleftarrow{S}_{41} & \overrightarrow{S}_{21}\,J_{32}\,J_{34}\overrightarrow{S}_{41} & J_{21}\overleftarrow{S}_{32}\,J_{34}\overleftarrow{S}_{41} \\
2J+R+S: & \overrightarrow{R}_{21}\,J_{32}\,J_{34}\overleftarrow{S}_{41} & \overleftarrow{R}_{21}\,J_{32}\,J_{34}\overrightarrow{S}_{41} & J_{21}\overleftarrow{S}_{32}\,J_{34}\overrightarrow{R}_{41}.
\end{array}$$

Acknowledgment. The author would like to thank Professor R. Jeltsch and Professor R. J. LeVeque for their helpful comments.

REFERENCES

[1] J. Glimm, C. Klingenberg, O. McBryan, B. Plohr, D. Sharp, and S. Yaniv, *Front tracking and two-dimensional Riemann problems*, Adv. in Appl. Math., 6 (1985), pp. 259–290.

[2] J. Guckenheimer, *Shocks and rarefactions in two space dimensions*, Arch. Rational Mech. Anal., 59 (1975), pp. 281–291.

[3] W. B. Lindquist, *The scalar Riemann problem in two spatial dimensions: piecewise smoothness of solutions and its breakdown*, SIAM J. Math. Anal., 17 (1986), pp. 1178–1197.

[4] ———, *Construction of solutions for two-dimensional Riemann problems*, Comput. Math. Appl., 12A (1986), pp. 615–630.

[5] C. W. Schulz-Rinne, J. P. Collins, and H. M. Glaz, *Numerical solution of the Riemann problem for two-dimensional gas dynamics*, Seminar für Angewandte Mathematik, ETH Zürich, Research Report No. 92-02, March 1992; SIAM. J. Sci. Comput. to appear.

[6] R. Smith, *The Riemann problem in gas dynamics*, Trans. Amer. Math. Soc., 249 (1979), pp. 1–50.

[7] J. A. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1982.

[8] D. H. Wagner, *The Riemann problem in two space dimensions for a single conservation law*, SIAM J. Math. Anal., 14 (1983), pp. 534–559.

[9] T. Chang (T. Zhang) and L. Hsiao (L. Xiao), *The Riemann Problem and Interaction of Waves in Gas Dynamics*, Pitman Monographs 41, Longman Scientific and Technical, Essex, 1989.

[10] T. Zhang and Y. Zheng, *Two-dimensional Riemann problem for a scalar conservation law*, Trans. Amer. Math. Soc., 312 (1989), pp. 589–619.

[11] ———, *Conjecture on the structure of solutions of the Riemann problem for two-dimensional gas dynamic systems*, SIAM J. Math. Anal., 21 (1990), pp. 593–630.

SIAM J. MATH. ANAL.
Vol. 24, No. 1, pp. 89-115, January 1993

007

THRESHOLD BEHAVIOR AND PROPAGATION FOR A DIFFERENTIAL-DIFFERENCE SYSTEM*

WEI-ZHENG GAO†

Abstract. This paper uses a comparison principle to study the long time behavior of a certain class of nonlinear difference-differential systems. These systems are motivated by certain models of heart tissue and myelinated axon. The main concern is to find steady state solutions, determine threshold properties, and study aspects of propagation. The theorems obtained are supported by numerical simulations.

Key words. threshold, comparison principle, myelinated axon, propagation

AMS(MOS) subject classifications. 34K25, 92A05

1. Introduction. Below we study a class of problems that are motivated by the structure of certain mathematical models, among them models of myelinated axon and heart tissue. The myelinated axon can be modeled by an electrical circuit. Based on current conservation from this circuit, we can set up the mathematical model to have the form

$$\frac{d}{dt}u_j = d(u_{j+1} - 2u_j + u_{j-1}) + f(u_j), \tag{1.1}$$

where u_j is the membrane potential at the jth node, $f(u)$ represents a current-voltage relation, and d is a constant [2], [3]. The simplest current-voltage relation is to let f be a cubic-shaped function. After appropriate scaling, we can assume

$$0 \leqq u_j \leqq 1 \quad \text{for all } j \in \mathbf{Z}.$$

Bell [2] studied this system. He proved that there exists a constant α, $0 < \alpha < 1$, such that if the initial condition $u_j(0) < \alpha$ holds for all j, then the solution $u_j(t) \to 0$ as $t \to \infty$ for all j. He calls this a subthreshold phenomenon. Under certain conditions, $u_j(t)$ does not go to zero as $t \to \infty$ for at least one j. This is called a superthreshold phenomenon. Under certain conditions on f, if $u_j(0)$ is greater than a particular value for at least one j, then $u_j(t) \to 1$ for all j as $t \to \infty$, and this process is in a form of propagating wave front. Keener [6] proved that propagation fails when the coupling is weak, i.e., when d is small. Zinner [8] proved that (1.1) has a traveling wavefront solution under certain conditions, and in [9] has shown that such a wavefront solution is globally stable.

To model a sheet of heart tissue, we need at least two spatial dimensions (Keener [7]). One possible mathematical model of a thin heart tissue is

$$\begin{aligned} \frac{d}{dt}u_{i,j} &= d_1(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) \\ &\quad + d_2(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) + f(u_{i,j}), \end{aligned} \tag{1.2}$$

where $u_{i,j}(t)$ represents the potential in the (i, j) cell at time t.

In this paper, we study the threshold behavior and propagation of a more general semidiscrete dynamic system

$$\frac{d}{dt}u_\alpha(t) - \underset{d}{\Delta} u_\alpha(t) - f_\alpha(u_\alpha) = 0,$$

* Received by the editors January 21, 1991; accepted for publication (in revised form) May 5, 1992.

† Department of Mathematical Sciences, Appalachian State University, Boone, North Carolina 28608.

where

$$\underset{d}{\Delta} u_\alpha \equiv \sum_{\beta \in C_\alpha} d_{\beta-\alpha}(u_\beta - u_\alpha) \quad \text{for } \alpha \in \mathbf{K} \subseteq \mathbf{Z}^d,$$

$\mathbf{K}$ and C_α are defined in § 2. This system includes (1.1) and (1.2) as special cases. Various results obtained here are parallel to Bell's results [2].

(1) There exists some constant q such that if $0 < u_\alpha(0) \leqq q$ for all α, then $u_\alpha(t) \to 0$ as $t \to \infty$ for all $\alpha \in \mathbf{Z}^d$ (Theorem 4.2). This is a subthreshold result.

(2) Under certain conditions on f_α, there exists a constant b, such that $u_\alpha(0) \geqq 0$ for all α and $u_\gamma(0) \geqq b$ for some γ guarantee that $\lim_{t\to\infty} u_\alpha = 1$ for all $\alpha \in \mathbf{Z}^d$ (Theorem 4.5). This is a superthreshold result.

(3) Under the conditions stated in (2), $u_\alpha \to 1$ in a form of a traveling wavefront. This behavior is studied in § 5. Estimation of the lower bound $\underline{c}$ and the upper bound $\bar{c}$ of the propagation speed are obtained (Theorems 5.1 and 5.2).

The above theoretical results are supported by numerical simulations that are discussed in § 6.

2. Model and comparison theorem. Let $\mathbf{K} \subseteq \mathbf{Z}^d$. $\mathbf{K} \neq \Phi$. We define the boundary of $\mathbf{K}$ as

$$\partial\mathbf{K} = \{\alpha \in \mathbf{Z}^d \mid \exists \beta \in \mathbf{K} \text{ and } \exists \gamma \notin \mathbf{K} \text{ such that } \|\alpha - \beta\| = 1 \text{ and } \|\alpha - \gamma\| = 1\}.$$

We also assume that $\operatorname{Int}\mathbf{K} = \mathbf{K} \backslash \partial\mathbf{K} \neq \Phi$, where $\operatorname{Int}\mathbf{K}$ is the interior of $\mathbf{K}$. Furthermore, for any $\alpha \in \mathbf{Z}^d$ we define

$$C_\alpha \equiv \{\beta \in \mathbf{Z}^d \,\|\, |\alpha_i - \beta_i| = 0 \text{ or } 1, i = 1, \ldots, d, \beta \neq \alpha\},$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$, $\beta = (\beta_1, \beta_2, \ldots, \beta_d)$. There are $N_d = 3^d - 1$ elements in C_α.

$$R^+ = \{t \mid 0 < t < \infty\},$$

$$R_0^+ = \{t \mid 0 \leqq t < \infty\},$$

$$X = \{\varphi \in C^1(R^+) \cap C(R_0^+) \mid \varphi'(0^+) \text{ exists}\}.$$

Consider the system

$$\frac{d}{dt} u_\alpha(t) - \underset{d}{\Delta} u_\alpha(t) - f_\alpha(u_\alpha) = 0, \tag{2.1}$$

where

$$\underset{d}{\Delta} u_\alpha \equiv \sum_{\beta \in C_\alpha} d_{\beta-\alpha}(u_\beta - u_\alpha) \quad \text{for } \alpha \in \mathbf{K}, \tag{2.2}$$

and the diffusion coefficients $d_\delta \geqq 0$, $\delta = \beta - \alpha$, with $\beta \in C_\alpha$, have the following properties:

(i) Symmetric with respect to α:

$$d_\delta = d_{-\delta}; \tag{2.3a}$$

(2.3b) (ii) $\sum_{\beta \in C_\alpha} d_\delta > 0;$

(2.3c) (iii) $d_\delta \equiv d_{\beta-\alpha}$ is independent of α.

Below, when we write $\sum_{\beta\in C_\alpha} d_\delta(u_\beta - u_\alpha)$, we will mean that $\delta = \beta - \alpha$. For a given dimension d, the operator Δ_d is uniquely determined by one particular set of d_δ's. We also suppose $f_\alpha \in C^1[0, 1]$ has the following properties:

(i) $f_\alpha(0) = f_\alpha(a_\alpha) = f_\alpha(1) = 0$, where $a_\alpha \in (0, 1)$,

(ii) $f_\alpha(u) < 0$ for $0 < u < a_\alpha$ and

(2.4) $f_\alpha(u) > 0$ for $a_\alpha < u < 1$,

where

$$0 < \underline{a} \leqq a_\alpha \leqq \bar{a} < 1, \quad \underline{a} = \inf_\alpha\{a_\alpha\}, \quad \bar{a} = \sup_\alpha\{a_\alpha\};$$

(iii) Assume

$$r_\alpha = \inf_s \left\{ 0 < s < a_\alpha \,\middle|\, f_\alpha(u) \leqq \frac{f_\alpha(p)}{p} u \quad \forall 0 < u \leqq p,\ s \leqq p \leqq a_\alpha \right\} < \underline{a}$$

for each α, $\bar{r} \equiv \sup_\alpha r_\alpha < \underline{a}$, and $\sup_{\alpha\in Z^d}\{f_\alpha(q)/q\} < 0$ for $q < \underline{a}$;

(iv) df_α/du is uniformly bounded over $0 \leqq u \leqq 1$.

The assumptions (i), (ii) on f_α may be dropped for some theorems and lemmas.

We first consider some examples of system (2.1).

Example 2.1. Special case of dimension $d = 1$.

Where α, β are scalar, let $\alpha \equiv i$; then $|\beta - \alpha| \leqq 1$ implies $\beta = i, i \pm 1$, so, $C_\alpha = \{i - 1, i + 1\}$. There are two diffusion coefficients, $d_{(i-1)-i} = d_{-1}$ and $d_{(i+1)-i} = d_1$. Equations (2.3a), (2.3b) imply that $d_{-1} = d_1 = d > 0$. Therefore, the operator $\Delta_d u_i$ is determined as

$$\begin{aligned} \underset{d}{\Delta} u_i &= d_1(u_{i+1} - u_i) + d_{-1}(u_{i-1} - u_i) \\ &= d(u_{i+1} - 2u_i + u_{i-1}), \end{aligned} \tag{2.5}$$

and (2.1) becomes

$$\frac{d}{dt} u_i = d(u_{i+1} - 2u_i + u_{i-1}) + f_i(u_i). \tag{2.6}$$

This is the prototypical example of system (2.1), and it has been studied extensively [2], [3].

Example 2.2. Special case of dimension $d = 2$.

In this case, $\mathbf{K} \subseteq \mathbf{Z}^2$; if $\alpha \equiv (i, j)$, then

$$C_{(i,j)} = \{(i \pm 1, j),\ (i, j \pm 1),\ (i \pm 1, j \pm 1)\}.$$

See Fig. 1(a). There are eight d_δ's: $d_{\pm 1,0}$, $d_{0,\pm 1}$, $d_{\pm 1,\pm 1}$. Different operators $\Delta_d u_{i,j}$ can be obtained by choosing different sets of d_δ's.

(1) For any $(i, j) \in \mathbf{Z}^2$, choosing

$$d_{1,0} = d_{-1,0} = d_1 > 0, \qquad d_{0,1} = d_{0,-1} = d_2 > 0,$$

$$d_{\pm 1,\pm 1} = 0; \quad \text{see Fig. 1(b)},$$

and with conditions (2.3a), (2.3b), and (2.3c) satisfied, we obtain

$$\underset{d}{\Delta} u_{i,j} = d_1(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + d_2(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}), \tag{2.7}$$

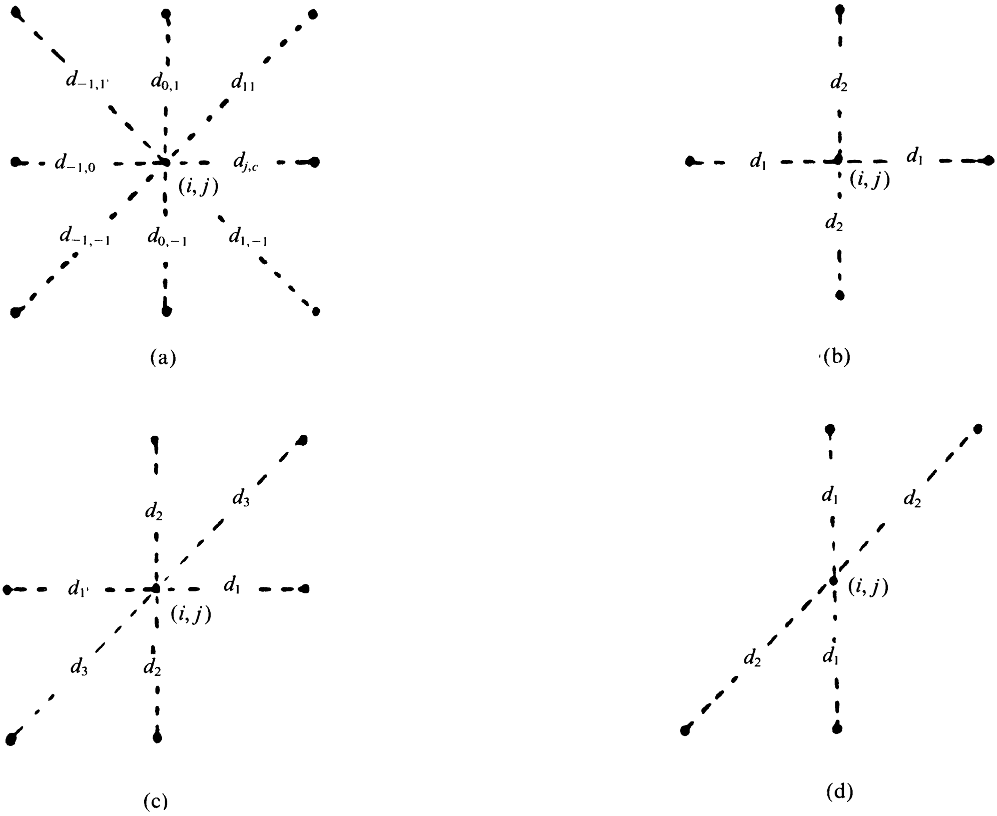


FIG 1. *Some examples of operator* $\underset{d}{\Delta}$ *for* $d=2$.

and (2.1) becomes

$$(2.8)\qquad \frac{d}{dt}u_{i,j}=d_1(u_{i+1,j}-2u_{i,j}+u_{i-1,j})+d_2(u_{i,j+1}-2u_{i,j}+u_{i,j-1})+f_{i,j}(u_{i,j}).$$

This is the standard two-dimensional analogue of (2.6).

(2) For any $(i,j)\in\mathbf{Z}^2$, choosing

$$d_{\pm1,0}=d_1>0,\quad d_{0,\pm1}=d_2>0,\quad d_{1,1}=d_{-1,-1}=d_3>0,$$
$$\text{and}\quad d_{1,-1}=d_{-1,1}=0$$

(see Fig. 1(c)), (2.3a) to (2.3c) are also satisfied, and (2.1) becomes

$$(2.9)\qquad \begin{aligned}\frac{d}{dt}u_{i,j}&=d_1(u_{i+1,j}-2u_{i,j}+u_{i-1,j})+d_2(u_{i,j+1}-2u_{i,j}+u_{i,j-1})\\&\quad+d_3(u_{i+1,j+1}-2u_{i,j}+u_{i-1,j-1})+f_{i,j}(u_{i,j}).\end{aligned}$$

(3) Similarly, if we choose

$$d_{0,\pm1}=d_1>0,\qquad d_{1,1}=d_{-1,-1}=d_2>0,$$
$$d_{\pm1,0}=0,\quad\text{and}\quad d_{1,-1}=d_{-1,1}=0$$

(see Fig. 1(d)), (2.1) becomes

$$\text{(2.10)}\qquad \frac{d}{dt}u_{i,j}=d_1(u_{i,j+1}-2u_{i,j}+u_{i,j-1})+d_2(u_{i+1,j+1}-2u_{i,j}+u_{i-1,j-1})+f_{i,j}(u_{i,j}).$$

Our main mathematical tool is the comparison theorem, which is stated as Lemma 2.2.

Let $\varphi_\alpha \in X$, f_α, $g_\alpha \in C^1(R^2)$; we define

$$\text{(2.11a)}\qquad P\varphi_\alpha(t)\equiv\frac{d}{dt}\varphi_\alpha(t)-\sum_{\beta\in C_\alpha} d_\delta(\varphi_\beta-k\varphi_\alpha)-f_\alpha(t,\varphi_\alpha),$$

$$\text{(2.11b)}\qquad \bar{P}\varphi_\alpha(t)\equiv\frac{d}{dt}\varphi_\alpha(t)-\sum_{\beta\in C_\alpha} d_\delta(\varphi_\beta-k\varphi_\alpha)-g_\alpha(t,\varphi_\alpha),$$

where $\alpha\in\mathbf{K}$, $k\in R$ is a constant, and the d_δ's satisfy (2.3).

LEMMA 2.1. *Assume* $\{u_\alpha(t)\}_{\alpha\in\mathbf{K}}$, $\{v_\alpha(t)\}_{\alpha\in\mathbf{K}}$ *with* u_α, $v_\alpha\in X$, f_α, $g_\alpha\in C^1(R^2)$, *and for all* $\alpha\in\mathbf{K}$ *the following conditions are satisfied*:

(2.12) (i) $Pu_\alpha \geqq \bar{P}v_\alpha$ *in* R^+, $\alpha\notin\partial\mathbf{K}$;

(2.13) (ii) $u_\alpha(0)\geqq v_\alpha(0)$;

(2.14) (iii) $f_\alpha(t,z)\geqq g_\alpha(t,z)$ *for* $t, z\in R$;

(iv) *If* $\mathbf{K}$ *has a boundary* $\partial\mathbf{K}$, *assume that*

$$\text{(2.15)}\qquad u_\gamma(t)\geqq v_\gamma(t)\quad\text{in } R_0^+\quad \forall\gamma\in\partial\mathbf{K};$$

(v) $\sup\{(\partial f_\alpha/\partial z)(t,z)\}=M<\infty$, *where the* sup *is taken over all* $z\in R$,

$$\alpha\in\mathbf{K},\quad\text{and}\quad t\geqq 0.$$

Then

$$u_\alpha(t)\geqq v_\alpha(t)\quad\text{in } R_0^+\quad\text{for all } \alpha\in\mathbf{K}.$$

Proof of this lemma is similar to the proof of Theorem 2.3 in [6]. The following comparison result is a special case of Lemma 2.1, where $k=1$, and $f_\alpha(t,z)=f_\alpha(z)$. This lemma is the main tool for studying system (2.1)-(2.4).

LEMMA 2.2. *Assume* $\{u_\alpha(t)\}_{\alpha\in\mathbf{K}}$, $\{v_\alpha(t)\}_{\alpha\in\mathbf{K}}$ *with* u_α, $v_\alpha\in X$, f_α, $g_\alpha\in C^1(R)$, *and for all* $\alpha\in\mathbf{K}$ *the following conditions are satisfied*:

(i) $\frac{d}{dt}u_\alpha-\underset{d}{\Delta}u_\alpha-f_\alpha(u_\alpha)\geqq\frac{d}{dt}v_\alpha-\underset{d}{\Delta}v_\alpha-g_\alpha(v_\alpha)$ *in* R^+, $\alpha\notin\partial\mathbf{K}$;

(ii) $f_\alpha(z)\geqq g_\alpha(z)$ *for* $z\in R$;

(iii) *For some* $M_1<\infty$, $\sup_{\alpha\in\mathbf{K}}\sup_{z\in R}\left\{\left|\frac{df_\alpha(z)}{dz}\right|\right\}\leqq M_1$,

(iv) $u_\alpha(0)\geqq v_\alpha(0)$,

(v) *Boundary condition* (2.15) *holds.*

Then

$$u_\alpha(t)\geqq v_\alpha(t)\quad\text{in } R_0^+\quad\text{for all } \alpha\in\mathbf{K}.$$

3. Steady state solutions. Before we go into the discussion of the threshold behavior, let us find some steady state solutions of (2.1)-(2.4) first, i.e., solutions to the system

$$\underset{d}{\Delta} u_\alpha + f_\alpha(u_\alpha) = 0,$$

or more explicitly,

$$\sum_{\beta \in C_\alpha} d_\delta(u_\beta - u_\alpha) + f_\alpha(u_\alpha) = 0. \tag{3.1}$$

3.1. Trivial steady state solutions. Trivial steady state solutions are $\{u_\alpha\} = \{0\}$, $\{1\}$ because $f_\alpha(0) = f_\alpha(1) = 0$ for all $\alpha \in \mathbf{Z}^d$. If $a_\alpha = a$ for all α, then $f_\alpha(a) = f_\alpha(a_\alpha) = 0$ for all α, which implies $\{u_\alpha\} = \{a\}$ is also a steady state solution.

We are interested in finding some nontrivial steady state solutions of system (2.1) with operator Δ_d given in (2.7). We consider the special case where $f_\alpha = f$ for all α, i.e.,

$$d_1(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + d_2(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) + f(u_{i,j}) = 0, \tag{3.2}$$

where $d_1 > 0$, $d_2 > 0$, and $f(u_{i,j})$ satisfies (2.4). Function $f(u_{i,j})$ is shown in Fig. 2.

3.2. A particular class of nontrivial symmetric solutions of (3.2). Let us see an example first.

Example. Let p, q be constants satisfying

$$0 < p < a < q < 1 \quad (\text{so } f(p) < 0,\ f(q) > 0),$$

and

$$\begin{aligned} &2(q-p)(d_1 + d_2) + f(p) = 0, \\ &2(p-q)(d_1 + d_2) + f(q) = 0. \end{aligned} \tag{3.3}$$

Let

$$\begin{aligned} &q_{0,0} = p, \\ &q_{i,j} = \begin{cases} p & \text{if } |i| + |j| \text{ is even}, \\ q & \text{if } |i| + |j| \text{ is odd}. \end{cases} \end{aligned} \tag{3.4}$$

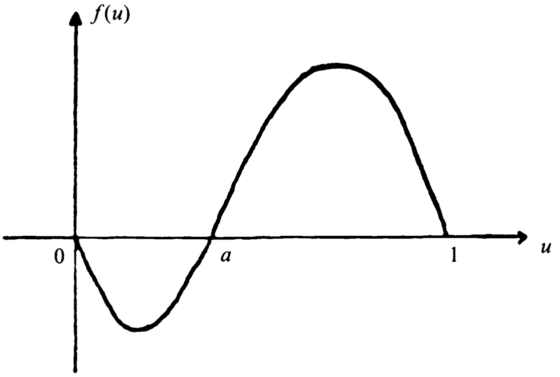


FIG. 2. *Typical form of $f(u)$.*

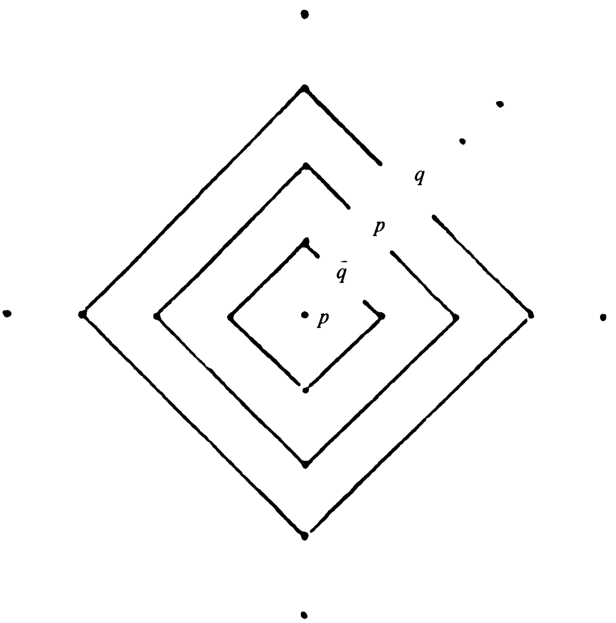


FIG. 3

Then $\{q_{i,j}\}$, defined by (3.4), is a (nontrivial) steady state solution of (2.1) and is given schematically in Fig. 3. This can be checked easily. For example, if for some $i, j \in \mathbf{Z}$ $|i|+|j|=$ even, then $|i\pm 1|+|j|$, $|i|+|j\pm 1|$ are both odd, so

$$q_{i,j}=p,\ q_{i+1,j}=q_{i-1,j}=q_{i,j+1}=q_{i,j-1}=q.$$

Therefore, by (3.3),

$$d_1(q_{i+1,j}-2q_{i,j}+q_{i-1,j})+d_2(q_{i,j+1}-2q_{i,j}+q_{i,j-1})+f(q_{i,j})=2(q-p)(d_1+d_2)+f(p)=0.$$

We now consider two particular cases.

(i) $f(u)=-(d_1+d_2)\sin(2\pi u)$, $d_1>0$, $d_2>0$ are arbitrary. f meets conditions of (2.4) with $a=\frac{1}{2}$. Let $p=\frac{1}{4}$, $q=\frac{3}{4}$, then

$$2(q-p)(d_1+d_2)+f(p)=(d_1+d_2)-(d_1+d_2)\sin\left(\frac{2\pi}{4}\right)=0,$$

$$2(p-q)(d_1+d_2)+f(q)=-(d_1+d_2)-(d_1+d_2)\sin\left(\frac{3\pi}{2}\right)=0,$$

so (3.3) is satisfied. From (3.4),

$$q_{i,j}=\begin{cases}\frac{1}{4} & \text{if } |i|+|j| \text{ is even (include 0)},\\ \frac{3}{4} & \text{if } |i|+|j| \text{ is odd},\end{cases}$$

so that $\{q_{i,j}\}$ is a nontrivial steady state solution.

(ii) $f(u)=0.5\cdot e^{5u}\sin[2\pi(1-u)^3]$, $\quad d_1=1, d_2=2.$

Here $a=1-2^{-1/3}\cong 0.2063$. This time p, q can be obtained by solving the nonlinear system that is represented in (3.3), namely,

$$6(q-p)+0.5\,e^{5p}\sin[2\pi(1-p)^3]=0,$$

$$6(p-q)+0.5\,e^{5q}\sin[2\pi(1-q)^3]=0.$$

The solution of this system is

$$p \cong 0.1117,$$

$$q \cong 0.2505.$$

Again, from (3.4) let

$$q_{i,j} = \begin{cases} p & \text{if } |i|+|j| \text{ is even (include 0),} \\ q & \text{if } |i|+|j| \text{ is odd,} \end{cases}$$

Then $\{q_{i,j}\}$ is a nontrivial steady state solution.

In the above example, the steady state $\{q_{i,j}\}$ is "diamond-symmetrically" distributed with respect to the "origin" $(i, j) = (0, 0)$, and attains values p and q alternatively. It is easy to prove that if the steady state solution $\{q_{i,j}\}$ has such a symmetric property and $d_1 \neq d_2$, then it must be obtained by the repeating of two values.

3.3. Some "special" steady state solutions of (2.1). There are some "special kinds" of steady state solutions of system (2.1). For example, let $\mathbf{K}$ be a bounded set centered at $\alpha = \mathbf{0}$, i.e., $\mathbf{K} = C_0 \cup \{\mathbf{0}\}$ (C_α is defined in § 2); therefore, (2.1) only holds at one point $\alpha = \mathbf{0}$, we can denote $f_\alpha = f$, and $a_\alpha = a$. Assume further that there exists b, such that

$$\sum_{\beta \in C_0} d_\beta b - f(b) = 0;$$

then $u_0(t) = b$, $u_\alpha(t) = 0$ for all $\alpha \in \partial\mathbf{K} \equiv C_0$ is a steady state solution. This follows from (3.1),

$$\begin{aligned} &\sum_{\beta \in C_0} d_\beta(u_\beta - u_0) + f(u_0) \\ &\quad = \sum_{\beta \in C_0} d_\beta(0 - u_0) + f(u_0) \\ &\quad = -\sum_{\beta \in C_0} d_\beta b + f(b) \\ &\quad = 0. \end{aligned}$$

This kind of steady state solution will be used when we discuss a superthreshold result.

4. Threshold results. We now investigate the sub- and superthreshold behavior of the system (2.1)-(2.4).

LEMMA 4.1. *Suppose* $\{u_\alpha(t)\}$, $\alpha \in \mathbf{Z}^d$ *is a solution to* (2.1)-(2.4) *with* $u_\alpha(0) \leqq q$ *for all* $\alpha \in \mathbf{Z}^d$. *If* $0 < q \leqq \underline{a}$, *then* $u_\alpha(t) \leqq q$ *for all* $\alpha \in \mathbf{Z}^d$, *and for all* $t \geqq 0$.

Proof. Let

$$w_\alpha(t) \equiv q \quad \forall \alpha \in \mathbf{Z}^d.$$

Then

$$\frac{d}{dt} w_\alpha - \underset{d}{\Delta} w_\alpha - f_\alpha(w_\alpha) = -f_\alpha(q) \geqq 0 = \frac{d}{dt} u_\alpha - \underset{d}{\Delta} u_\alpha - f_\alpha(u_\alpha)$$

$$w_\alpha(0) = q \geqq u_\alpha(0) \quad \forall \alpha.$$

By Lemma 2.2 $u_\alpha(t) \leqq w_\alpha(t) \equiv q$, for all $\alpha \in \mathbf{Z}^d$ and for all $t \geqq 0$. □

THEOREM 4.2. *Suppose* $\{u_\alpha(t)\}$ *is a solution to* (2.1)-(2.4) *with* $0 \leqq u_\alpha(0) \leqq q$ *for all* $\alpha \in \mathbf{Z}^d$. *If* $0 \leqq q < \underline{a}$, *then* $\lim_{t\to\infty} u_\alpha(t) = 0$ *uniformly for* $\alpha \in \mathbf{Z}^d$.

Proof. If $q = 0$, then $u_\alpha(t) \equiv 0$ for all α because zero is a steady state solution. We now assume $q > 0$, and $\bar{r} \leqq q < \underline{a}$ ($\bar{r}$ is defined in (2.4)). Let

$$\eta = \sup_{\alpha \in \mathbf{Z}^d} \left\{ \frac{f_\alpha(q)}{q} \right\} < 0 \quad ((\text{iii}) \text{ in } (2.4)), \tag{4.1}$$

$$w_\alpha(t) \equiv e^{\eta t}; \tag{4.2}$$

then $w_\alpha(0) = 1$ and

$$\begin{aligned} &\frac{d}{dt} w_\alpha - \underset{d}{\Delta} w_\alpha - \eta w_\alpha \\ &= \frac{d}{dt} w_\alpha - \sum_{\beta \in C_\alpha} d_\delta (w_\beta - w_\alpha) - \eta w_\alpha \\ &= \eta\, e^{\eta t} - 0 - \eta\, e^{\eta t} \\ &= 0. \end{aligned}$$

Suppose $u_\alpha(t)$ solves (2.1)-(2.4) with $0 \leqq u_\alpha(0) \leqq q$ for all α. Then,

$$\frac{d}{dt} u_\alpha - \underset{d}{\Delta} u_\alpha - f_\alpha(u_\alpha) = 0 = \frac{d}{dt} w_\alpha - \underset{d}{\Delta} w_\alpha - \eta w_\alpha,$$

$$0 \leqq u_\alpha(0) \leqq q \leqq 1 = w_\alpha(0).$$

Furthermore, $u_\alpha(0) \leqq q$ for all α, implies, by Lemma 4.1, $u_\alpha(t) \leqq q$ for all α, and for all $t \geqq 0$. For such $u_\alpha(t)$

$$f_\alpha(u_\alpha) \leqq \frac{f_\alpha(q)}{q} u_\alpha \leqq \eta u_\alpha.$$

The first inequality comes from the assumption that $q \geqq \bar{r}$. Therefore, the conditions of Lemma 2.2 are satisfied, so

$$0 \leqq u_\alpha(t) \leqq w_\alpha(t) \equiv e^{\eta t} \quad \forall t \geqq 0 \quad \text{and} \quad \forall \alpha \in \mathbf{Z}^d.$$

Finally, $\lim_{t\to\infty} e^{\eta t} = 0$ (since $\eta < 0$) implies $\lim_{t\to\infty} u_\alpha(t) = 0$ for all $\alpha \in \mathbf{Z}^d$.

If $q < \bar{r}$, then, let $v_\alpha(t)$ be the solution of (2.1)-(2.4) with $\bar{r} \leqq v_\alpha(0) < \underline{a} - \varepsilon$ for some $\varepsilon > 0$. By Lemma 2.2, $u_\alpha(t) \leqq v_\alpha(t)$. But from the above discussion, $v_\alpha(t) \to 0$ as $t \to \infty$, for all α, therefore, again we have $\lim_{t\to\infty} u_\alpha(t) = 0$ for all $\alpha \in \mathbf{Z}^d$. This completes the proof. □

The following corollary gives us a superthreshold result.

COROLLARY 4.3. *Suppose* $\{u_\alpha(t)\}$ *is a solution to* (2.1)-(2.4) *with* $p \leqq u_\alpha(0) \leqq 1$ *for all* $\alpha \in \mathbf{Z}^d$. *If* $p > \bar{a}$, *then* $u_\alpha(t) \to 1$ *as* $t \to \infty$ *for all* $\alpha \in \mathbf{Z}^d$.

Proof. This is a direct result of Theorem 4.2. Suppose $\{u_\alpha(t)\}$ satisfies the above conditions. Define $w_\alpha(t) \equiv 1 - u_\alpha(t)$, $a_\alpha^* \equiv 1 - a_\alpha$, $q \equiv 1 - p$, and $F_\alpha(w) \equiv -f_\alpha(1-w)$; then $F_\alpha(w)$ satisfies (2.4) with a_α replaced by a_α^*. It can be checked easily. For example, $F_\alpha(0) = -f_\alpha(1-0) = -f_\alpha(1) = 0$, and $F_\alpha(a_\alpha^*) = F_\alpha(1 - a_\alpha) = -f_\alpha(a_\alpha) = 0$. Further, if $w \in (0, a_\alpha^*)$, then $0 < w < a_\alpha^* \equiv 1 - a_\alpha$, which implies $a_\alpha < 1 - w < 1$, so $F_\alpha(w) = -f_\alpha(1-w) < 0$. Similarly, it can be shown that $F_\alpha(w) > 0$ for $w \in (a_\alpha^*, 1)$. Furthermore, let $\underline{a}^* = \inf a_\alpha^*$;

then $\underline{a}^* = \inf(1-a_\alpha) = 1 - \sup a_\alpha = 1 - \bar{a}$. We also have

$$\begin{aligned}
&\frac{d}{dt} w_\alpha - \underset{d}{\Delta} w_\alpha - F_\alpha(w_\alpha)\\
&= \left[-\frac{d}{dt} u_\alpha \right] - [-\underset{d}{\Delta} u_\alpha] - [-f_\alpha(1-w_\alpha)]\\
&= -\left\{ \frac{d}{dt} u_\alpha - \underset{d}{\Delta} u_\alpha - f_\alpha(u_\alpha) \right\}\\
&= 0
\end{aligned}$$

and

$$w_\alpha(0) = 1 - u_\alpha(0) \leqq 1 - p < 1 - \bar{a} = \underline{a}^* \quad \text{for all } \alpha \in \mathbf{Z}^d.$$

Thus $\{w_\alpha(t)\}$ satisfies the conditions of Theorem 4.2 with $\underline{a}$ replaced by $\underline{a}^*$, by the theorem, $w_\alpha(t) \to 0$ as $t \to \infty$ for all $\alpha \in \mathbf{Z}^d$. The final result is obtained by considering $w_\alpha(t) = 1 - u_\alpha(t)$. □

Theorem 4.2 gives a subthreshold result. We now look for conditions that guarantee that $u_\alpha(t) \to 1$ as $t \to \infty$ for all $\alpha \in \mathbf{Z}^d$. A trivial condition is given in Corollary 4.3, that is assuming $u_\alpha(0) > \bar{a}$ for all $\alpha \in \mathbf{Z}^d$. Theorem 4.5 below tells us that under certain conditions, $u_\alpha(0) > b > a$ holding only at one node is enough to bump up the entire solution $\{u_\alpha(t)\}$ to $\{1\}$.

The steady state solution of (2.1) in the statement of Lemma 4.4 exists under certain conditions on f_α and **K** as discussed at the end of § 3. This lemma is a straight generalization of Lemma 2 in Bell and Cosner [2] and is similar to Proposition 2.2 in Aronson and Weinberger [1].

LEMMA 4.4. *Suppose for some domain* $\mathbf{K} \subset \mathbf{Z}^d$, *and* $\alpha \in \mathbf{K}$, $\{q_\alpha\}$ *satisfies* $0 \leqq q_\alpha \leqq 1$ *and*

$$\underset{d}{\Delta} q_\alpha + f_\alpha(q_\alpha) = 0. \tag{4.3}$$

If **K** *has boundary* $\partial\mathbf{K}$, *assume further* $q_\gamma \leqq 0$ *for all* $\gamma \in \partial\mathbf{K}$. *Let* $\{u_\alpha(t)\}$ *be a solution of* (2.1) *with* $u_\alpha(0) = q_\alpha$ *for* $\alpha \in \mathbf{K}$ *and* $u_\alpha(0) = 0$ *for all* $\alpha \in \mathbf{Z}^d \backslash \mathbf{K}$. *Then for each* α, $u_\alpha(t)$ *is a nondecreasing function of t with*

$$\lim_{t \to \infty} u_\alpha(t) = \tau_\alpha,$$

where $\{\tau_\alpha\}$ *is the smallest nonnegative solution to* (4.3) *valid for all* $\alpha \in \mathbf{Z}^d$, *which satisfies* $\tau_\alpha \geqq q_\alpha$ *for* $\alpha \in \mathbf{K}$.

In Theorem 4.5, we assume that $f_\alpha(u) = f(u)$ for all $\alpha \in \mathbf{Z}^d$, where f satisfies (2.4) with $a_\alpha = a$ fixed.

THEOREM 4.5. *Suppose* $f(u)$ *satisfies* (2.4). *Let* $M \equiv \sup_{0 \leqq u \leqq 1} f(u) > 0$, $m \equiv -\inf_{0 \leqq u \leqq 1} f(u) > 0$, *and suppose there exists constants* b, e *with* $a < b < e < 1$ *such that* (*see Fig.* 4)

$$\begin{aligned}
&2\left(\sum_{i=1}^{s} d_i \right) u - f(u) < 0 \quad \textit{for } u \in (b, e),\\
&2\left(\sum_{i=1}^{s} d_i \right) u - f(u) = 0 \quad \textit{at } u = b, e,
\end{aligned} \tag{4.4}$$

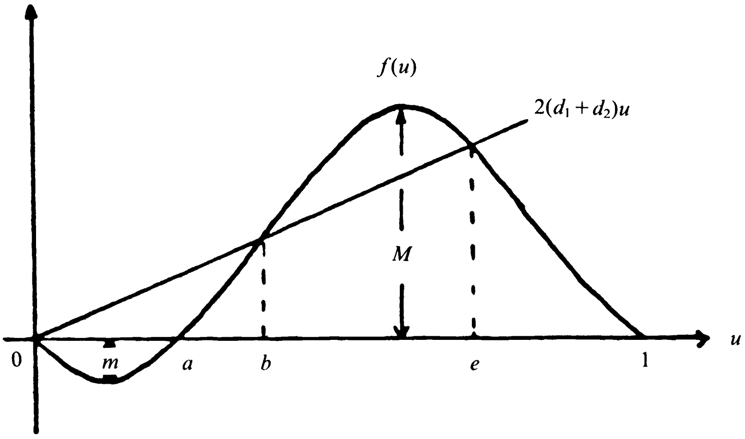


FIG. 4

where $d_i > 0$ for $i = 1, 2, \dots, 2s$, and s is to be defined such that $2s$ is the number of d_δ's that are >0. Suppose one of the following holds:

$$2\left(\sum_{i=1}^{s} d_i\right)b + m \leqq e\cdot\min(d_1, d_2, \dots, d_s) \tag{4.5}$$

$$2\left(\sum_{i=1}^{s} d_i\right) + M - (1-b)\min(d_1, d_2, \dots, d_s) \leqq 2\left(\sum_{i=1}^{s} d_i\right)e. \tag{4.6}$$

If $\{u_\alpha(t)\}$ is a solution of (2.1) *with $u_\alpha(0) \geqq 0$ for all α and $u_\gamma(0) \geqq b$ for some γ, then* $\lim_{t\to\infty} u_\alpha(t) = 1$ *for all $\alpha \in \mathbf{Z}^d$.*

The proof of the theorem is based on the following four lemmas. In order to reduce the bookkeeping work as much as possible, those lemmas are stated in the general form but proved in detail for $d = 2$, and for the particular operator $\Delta_d u_{i,j}$ given in (2.7):

$$\underset{d}{\Delta} u_{i,j} = d_1(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + d_2(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}).$$

Throughout the next four lemmas, we assume that $f(u)$ satisfies the conditions in Theorem 4.5, (4.4) holds, and $\{q_\alpha\}$ is the smallest global nonnegative steady state solution to (2.1)-(2.4) with $q_\alpha \geqq b$ for some α.

LEMMA 4.6. *$q_\alpha \notin [b, e]$ for all $\alpha \in \mathbf{Z}^d$.*

Proof. Suppose there exists $\alpha = (i, j)$ with $b \leqq q_{i,j} \leqq e$. By (3.1),

$$\underset{d}{\Delta} q_{i,j} + f(q_{i,j}) = 0,$$

i.e.,

$$d_1(q_{i+1,j} - 2q_{i,j} + q_{i-1,j}) + d_2(q_{i,j+1} - 2q_{i,j} + q_{i,j-1}) + f(q_{i,j}) = 0,$$

or

$$\begin{aligned} d_1(q_{i+1,j} + q_{i-1,j}) + d_2(q_{i,j+1} + q_{i,j-1}) &= 2(d_1+d_2)q_{i,j} - f(q_{i,j}) \\ &\leqq 0 \quad (\text{by } (4.4)). \end{aligned}$$

So at least one of $q_{i+1,j}$, $q_{i-1,j}$, $q_{i,j+1}$, $q_{i,j-1}$ must be $\leqq 0$. For example, let $q_{i+1,j} \leqq 0$. If $q_{i+1,j} < 0$, then $\{q_{i,j}\}$ is not a global nonnegative steady state solution, with $q_{i+1,j} = 0$; we obtain,

$$d_1(q_{i+2,j} - 2q_{i+1,j} + q_{i,j}) + d_2(q_{i+1,j+1} - 2q_{i+1,j} + q_{i+1,j-1}) + f(q_{i+1,j}) = 0,$$

or

$$d_1 q_{i+2,j} + d_2(q_{i+1,j+1} + q_{i+1,j-1}) = -d_1 q_{i,j} < 0$$

$$(q_{i,j} > 0 \text{ because } q_{i,j} \in [b, e]).$$

Therefore, at least one of $q_{i+1,j}$, $q_{i-1,j}$, $q_{i,j+1}$, $q_{i,j-1}$ must be negative, so $\{q_{i,j}\}$ is not a nonnegative steady state. □

LEMMA 4.7. *Assume further that* (4.5) *holds. If there exists* $\alpha_\circ \in \mathbf{Z}^d$ *with* $q_{\alpha_\circ} \leqq b$, *then* $q_\alpha < b$ *for all* $\alpha \in \mathbf{Z}^d$.

Proof. Without loss of generality, let us assume $q_{0,0} < b$. From (3.1) and (4.5),

$$\begin{aligned} d_1(q_{1,0} + q_{-1,0}) + d_2(q_{0,1} + q_{0,-1}) \\ = 2(d_1 + d_2)q_{0,0} - f(q_{0,0}) \\ \leqq 2(d_1 + d_2)b + m \leqq e \cdot \min(d_1, d_2), \end{aligned}$$

i.e.,

$$d_1 q_{1,0} + d_1 q_{-1,0} + d_2 q_{0,1} + d_2 q_{0,-1} \leqq e \cdot \min(d_1, d_2).$$

So every term on the left-hand side is $\leqq e \cdot \min(d_1, d_2)$ because all d_1, d_2 and q's are nonnegative. But, $d_1 q_{1,0} \leqq e \cdot \min(d_1, d_2)$ implies $q_{1,0} \leqq e \cdot (\min(d_1, d_2)/d_1) \leqq e$. Similarly, $q_{-1,0} \leqq e$, $q_{0,1} \leqq e$, and $q_{0,-1} \leqq e$. Furthermore, by Lemma 4.6, $q_{i,j} \notin [b, e]$ implies $q_{1,0} < b$, $q_{-1,0} < b$, $q_{0,1} < b$, and $q_{0,-1} < b$.

It follows by induction that $q_{i,j} < b$ for all $i, j \in \mathbf{Z}$. □

LEMMA 4.8. *Assume further that* (4.6) *holds. If there exists* $\alpha_\circ \in \mathbf{Z}^d$ *with* $q_{\alpha_\circ} > e$, *then* $q_\alpha > e$ *for all* $\alpha \in \mathbf{Z}^d$.

Proof. Let us assume $q_{0,0} > e$. Define

$$\bar{q}_{i,j} \equiv 1 - q_{i,j}, \qquad F(\bar{q}_{i,j}) \equiv -f(1 - \bar{q}_{i,j});$$

then (3.1) can be rewritten as

$$d_1(\bar{q}_{i+1,j} - 2\bar{q}_{i,j} + \bar{q}_{i-1,j}) + d_2(\bar{q}_{i,j+1} - 2\bar{q}_{i,j} + \bar{q}_{i,j-1}) + F(\bar{q}_{i,j}) = 0. \tag{4.7}$$

We have

$$\sup_{0 \leqq u \leqq 1} (-F(u)) = \sup_{0 \leqq u \leqq 1} (f(1-u)) = M.$$

Notice that $q_{0,0} > e$ implies $\bar{q}_{0,0} < 1 - e$, and from (4.7), (4.6),

$$\begin{aligned} d_1\bar{q}_{1,0} + d_1\bar{q}_{-1,0} + d_2\bar{q}_{0,1} + d_2\bar{q}_{0,-1} &= 2(d_1 + d_2)\bar{q}_{0,0} - F(\bar{q}_{0,0}) \\ &< 2(d_1 + d_2)(1 - e) + M \\ &\leqq (1 - b) \cdot \min(d_1, d_2). \end{aligned}$$

Therefore, every term on the left-hand side is $<1-(b)\cdot\min(d_1, d_2)$. But $d_1\bar{q}_{1,0}<(1-b)\min(d_1, d_2)$ implies $\bar{q}_{1,0}<(1-b)[\min(d_1,d_2)/d_1]\leqq 1-b$, i.e., $q_{1,0}>b$, but $q_{i,j}\notin[b,e]$ by Lemma 4.6, thus $q_{1,0}>e$. Similarly, $q_{-1,0}>e$, $q_{0,1}>e$, and $q_{0,-1}>e$. By induction, $q_{i,j}>e$ for all $i,j\in\mathbf{Z}$. □

LEMMA 4.9. *Assume further that* (4.5) *or* (4.6) *holds; then* $q_\alpha\notin(e,1)$ *for all* $\alpha\in\mathbf{Z}^d$.

Proof. Suppose $q_{0,0}\in(e,1)$. First, let us construct a sequence $\{q_k\}$ out of $\{q_{i,j}\}$ such that $q_{0,0}\equiv q_0>q_1>q_2>\cdots>q_m>q_{m+1}$ with $q_m>e>a$ and $q_{m+1}\leqq e$, where $q_k\in\{q_{i,j}\}$ for $k=0,1,\ldots,m+1$ as follows.

Let $(i,j)=(0,0)$ in (3.1); define $q_0\equiv q_{0,0}$. We obtain

$$d_1(q_{1,0}-2q_{0,0}+q_{-1,0})+d_2(q_{0,1}-2q_{0,0}+q_{0,-1})+f(q_{0,0})=0$$

or

$$d_1(q_{1,0}-q_0)+d_1(q_{-1,0}-q_0)+d_2(q_{0,1}-q_0)+d_2(q_{0,-1}-q_0)=-f(q_0).$$

Assume without loss of generality that

$$q_{1,0}-q_0=\min(q_{1,0}-q_0, q_{-1,0}-q_0, q_{0,1}-q_0, q_{0,-1}-q_0).$$

Define $q_1\equiv q_{1,0}$ and

$$\varepsilon=\frac{1}{2(d_1+d_2)}\cdot\inf_{q\in[e,q_0]}\{f(q)\}>0.$$

We obtain

$$\begin{aligned}&2(d_1+d_2)(q_1-q_0)\\&\quad\leqq d_1(q_{1,0}-q_0)+d_1(q_{-1,0}-q_0)+d_2(q_{0,1}-q_0)+d_2(q_{0,-1}-q_0)\\&\quad=-f(q_0);\end{aligned}$$

therefore,

$$q_1-q_0\leqq\frac{-f(q_0)}{2(d_1+d_2)}\leqq-\varepsilon.$$

If $q_1\leqq e$, we are done ($m=0$).

Suppose $q_1>e$. Let $(i,j)=(1,0)$ in (3.1); we have

$$d_1(q_{2,0}-q_1)+d_1(q_{0,0}-q_1)+d_2(q_{1,1}-q_1)+d_2(q_{1,-1}-q_1)=-f(q_{1,0}).$$

Let q_2 be one of $q_{2,0}$, $q_{0,0}$, $q_{1,1}$, and $q_{1,-1}$, such that

$$q_2-q_1=\min(q_{2,0}-q_1, q_{0,0}-q_1, q_{1,1}-q_1, q_{1,-1}-q_1);$$

then

$$\begin{aligned}&2(d_1+d_2)(q_2-q_1)\\&\quad\leqq d_1(q_{2,0}-q_1)+d_1(q_{0,0}-q_1)+d_2(q_{1,1}-q_1)+d_2(q_{1,-1}-q_1)\\&\quad=-f(q_1);\end{aligned}$$

therefore,

$$q_2-q_1\leqq\frac{-f(q_1)}{2(d_1+d_2)}\leqq-\varepsilon.$$

Following the same procedure, we can construct a sequence $q_{0,0} \equiv q_0 > q_1 > q_2 > \cdots > q_k > q_{k+1}$, where

$$q_{k+1} - q_k \leqq -\varepsilon \quad \text{if } e < q_k \leqq q_0.$$

Therefore,

$$q_{k+1} \leqq q_k - \varepsilon \leqq q_{k-1} - 2\varepsilon \leqq \cdots \leqq q_0 - (k+1)\varepsilon.$$

This equation holds as long as $q_k > e$. It is easy to see that when k increases, q_{k+1} will eventually drop below e or be equal to it. This completes our construction of the sequence $\{q_k\}$.

The assumption we made so far is only (4.4). Now consider (4.5) and (4.6).

Case 1. Assume (4.5) holds. According to Lemma 4.6, $q_{m+1} \leqq e$ implies $q_{m+1} < b$. Therefore, $q_{i,j} < b$ for all i, j, by Lemma 4.7, which contradicts the assumption $q_{0,0} \in (e, 1)$.

Case 2. Assume (4.6) holds. By Lemma 4.8, $q_{0,0} > e$ implies $q_{i,j} > e$ for all i, j, which contradicts $q_{m+1} \leqq e$.

The contradictions are caused by the assumption $q_{0,0} \in (e, 1)$; therefore, $q_{i,j} \notin (e, 1)$ for all i, j. □

Now we can prove Theorem 4.5.

Proof of Theorem 4.5. First, from Lemmas 4.6 and 4.9, we conclude that the global nonnegative steady state solution $\{q_\alpha\}$ with $q_\alpha \geqq b$ for some α, under the assumptions of the theorem, must satisfy $q_\alpha \geqq 1$. But $q_\alpha \equiv 1$ is a steady state; therefore, the smallest such steady state is $q_\alpha \equiv 1$.

Let $\mathbf{0}$ be the d-dimensional zero vector. Suppose now $u_\alpha(0) \geqq 0$, and $u_{\mathbf{0}}(0) \geqq b$ (if $u_\alpha(0) \geqq b$ for some α, then simply make a translation in α). Let $\{v_\alpha(t)\}$ be the solution of (2.1) with initial data $v_\alpha(0) = 0$ for $\alpha \neq \mathbf{0}$ and $v_{\mathbf{0}} = b$. Since $q_\beta = 0$ for $\beta \in C_{\mathbf{0}} \backslash \{\mathbf{0}\}$, and $q_{\mathbf{0}} = b$ yields a solution of (4.3) at $\alpha = \mathbf{0}$ (where $2(\sum_{i=1}^{s} d_i)b - f(b) = 0$), we may apply Lemma 4.4 and conclude that for any α, $v_\alpha(t)$ is a monotone nondecreasing function of t. Hence $\lim_{t\to\infty} v_\alpha(t) = \tau_\alpha$, where $\{\tau_\alpha\}$ is the smallest global steady state for (2.1) with values lying above b. However, the only such steady state is $\{\tau_\alpha\} \equiv \{1\}$, so $\lim_{t\to\infty} v_\alpha(t) = 1$. Finally, $v_\alpha(0) \leqq u_\alpha(0) \leqq 1$ implies, by Lemma 2.2, $v_\alpha(t) \leqq u_\alpha(t) \leqq 1$ for all t, and all $\alpha \in \mathbf{Z}^d$; therefore, we obtain the conclusion that $\lim_{t\to\infty} u_\alpha(t) = 1$ for all $\alpha \in \mathbf{Z}^d$. □

THEOREM 4.10. *Suppose there are values* x_1, x_2, x_3, x_4 *with* $0 < x_1 < x_2 < x_3 < x_4 < 1$, *such that*

$$\text{(4.8)} \quad \text{(i)} \quad (1-x_1) + \frac{f(x_1)}{D} = 0, \qquad (1-x_2) + \frac{f(x_2)}{D} = 0, \quad \textit{and}$$
$$\frac{f(z)}{D} + (1-z) < 0 \quad \textit{for } 0 < x_1 < z < x_2,$$

$$\text{(4.9)} \quad \text{(ii)} \quad x_3 - \frac{f(x_3)}{D} = 0, \qquad x_4 - \frac{f(x_4)}{D} = 0, \quad \textit{and}$$
$$\frac{f(z)}{D} - z > 0 \quad \textit{for } x_3 < z < x_4 < 1,$$

where $D = \sum_{\beta \in C_\alpha} d_\delta$ (*independent of* α *by* (2.3c)). *See Fig.* 5. *Suppose* $0 \leqq u_\alpha(0) \leqq 1$ *for all* $\alpha \in \mathbf{Z}^d$. *Then for any* $\beta \in \mathbf{Z}^d$, *if* $u_\beta(0) \in [0, x_2)$, *then* $u_\beta(t) \in [0, x_2)$ *for all* $t > 0$. *If* $u_\beta(0) \in (x_3, 1]$, *then* $u_\beta(t) \in (x_3, 1]$ *for all* $t > 0$.

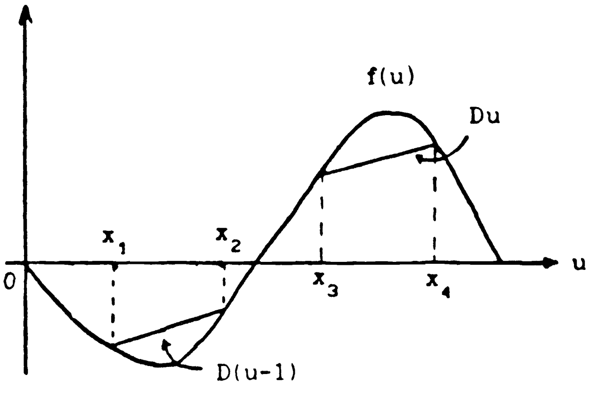


FIG. 5

Proof. First, $0 \leqq u_\alpha(0) \leqq 1$ for all $\alpha \in \mathbf{Z}^d$ implies $0 \leqq u_\alpha(t) \leqq 1$ for all $\alpha \in \mathbf{Z}^d$ and all $t \geqq 0$ by Lemma 2.2. Now suppose $u_\gamma(0) \in [0, x_2)$. If we can prove that whenever $u_\gamma(t)$ goes into region (x_1, x_2) it stops increasing, then $u_\gamma(t)$ will stay in the region $[0, x_2)$ for all t. This is true because if $u_\gamma(t) \in (x_1, x_2)$, then

$$\begin{aligned}\frac{d}{dt} u_\gamma(t) &= \sum_{\beta \in C_\gamma} d_\delta(u_\beta - u_\gamma) + f(u_\gamma) \\ &\leqq \sum_{\beta \in C_\gamma} d_\delta(1 - u_\gamma) + f(u_\gamma) \quad (\text{since } u_\beta \leqq 1 \text{ for all } \beta) \\ &= \left(\sum_{\beta \in C_\gamma} d_\delta\right)(1 - u_\gamma) + f(u_\gamma) \\ &< 0,\end{aligned}$$

by (4.8)).

The other part of the theorem can be proved in the analogous way.

5. Propagation. Throughout this section, we assume

$$\lim_{t \to \infty} u_\alpha(t) = 1 \quad \text{for all } \alpha \in \mathbf{Z}^d. \tag{5.1}$$

From Theorem 4.5 this is possible for some classes of f's. Below we use the notation

$$|\alpha| \equiv |(\alpha_1, \alpha_2, \ldots, \alpha_d)| = |\alpha_1| + |\alpha_2| + \cdots + |\alpha_d|,$$

and $[\cdot]$ means "integer part of."

THEOREM 5.1. *Assume*

(i) $\{u_\alpha(t)\}$ *is a solution to* (2.1)-(2.4) *satisfying* (5.1);

(ii) $u_\alpha(0) \in [0, 1]$ *for all* $\alpha \in \mathbf{Z}^d$,

$$u_\alpha(0) = 0 \quad \textit{for } |\alpha| > N \quad \textit{where } N < \infty. \tag{5.2}$$

Then there exists a $\bar{c} > 0$ *such that for each* $\alpha_0 \in \mathbf{Z}^d$, $\lim_{t \to \infty} u_\alpha(t) = 0$ *if* $|\alpha| = |\alpha_0| + [ct]$ *for* $c > \bar{c}$.

Proof. Define

$$\sigma = \sup_{\alpha \in \mathbf{Z}^d} \sup_{0 < u < 1} \left\{\frac{f_\alpha(u)}{u}\right\} > 0,$$

where $\sigma<\infty$ because of (2.4). Then

$$\begin{aligned}
\Lambda u_\alpha &\equiv \underset{d}{\Delta} u_\alpha+\sigma u_{\alpha'} \qquad (5.3)\\
&= \sum_{\beta\in C_\alpha} d_\delta(u_\beta-u_\alpha)+\sum_{\beta\in C_\alpha}\left(\frac{\sigma}{\sum_{\beta'\in C_\alpha} d_{\delta'}} d_\delta u_{\alpha'}\right)(\delta=\beta-\alpha,\ \delta'=\beta'-\alpha)\\
&= \sum_{\beta\in C_\alpha} d_\delta\left[u_\beta-\left(1-\frac{\sigma}{\sum_{\beta'\in C_\alpha} d_{\delta'}}\right)u_\alpha\right]\\
&= \sum_{\beta\in C_\alpha} d_\delta(u_\beta-ku_\alpha), \qquad (5.4)
\end{aligned}$$

where

$$k\equiv 1-\frac{\sigma}{\sum_{\beta'\in C_\alpha} d_{\delta'}}. \tag{5.5}$$

Equation (2.1) may be rewritten as

$$\frac{d}{dt}u_\alpha-\underset{d}{\Delta} u_\alpha-f_\alpha(u_\alpha)=\frac{d}{dt}u_\alpha-\Lambda u_\alpha+\sigma u_\alpha-f_\alpha(u_\alpha)=0,$$

or

$$\frac{d}{dt}u_\alpha-\Lambda u_\alpha=f_\alpha(u_\alpha)-\sigma u_\alpha\leqq 0. \tag{5.6}$$

The last inequality is obtained by using the fact that $u_\alpha(0)\in[0,1]$ implies, by Lemma 2.2, $u_\alpha(t)\in[0,1]$, and by using the definition of σ.

Define

$$w_\alpha(t)=Ap(t)\,e^{-\mu(|\alpha|-ct)}, \tag{5.7}$$

where positive constants μ, c, and A, and function $p(t)\geqq 0$ are to be determined. Then, by (5.4),

$$\begin{aligned}
\Lambda w_\alpha &= \sum_{\beta\in C_\alpha} d_\delta(w_\beta-kw_\alpha)\\
&= \sum_{\beta\in C_\alpha} d_\delta[Ap(t)\,e^{-\mu(|\beta|-ct)}-kAp(t)\,e^{-\mu(|\alpha|-ct)}]\\
&= Ap(t)\,e^{-\mu(|\alpha|-ct)}\sum_{\beta\in C_\alpha} d_\delta[e^{-\mu(|\beta|-|\alpha|)}-k] \qquad (5.8)\\
&= Ap(t)\,e^{-\mu(|\alpha|-ct)}\Gamma_\alpha(\mu|\underset{d}{\Delta}),
\end{aligned}$$

where

$$\Gamma_\alpha(\mu|\underset{d}{\Delta})\equiv\sum_{\beta\in C_\alpha} d_\delta[e^{-\mu(|\beta|-|\alpha|)}-k],\qquad \delta=\beta-\alpha. \tag{5.9}$$

$\Gamma_\alpha(\mu|\Delta_d)$ is a function of μ depending on α and the particular form of operator Δ_d. Explicit representations of $\Gamma_\alpha(\mu|\Delta_d)$ for a couple of particular Δ_d's are given in Examples 5.1 and 5.2. Furthermore, considering there are 3^d-1 elements in C_α and the definition of C_α, it is true that for any fixed μ, $-\mu(|\beta|-|\alpha|)$ is bounded above; i.e., there exists $M<\infty$ such that $-\mu(|\beta|-|\alpha|)<\mu M$. Therefore,

$$\Gamma_\alpha(\mu|\underset{d}{\Delta})\leqq[e^{\mu M}-k]\cdot\sum_{\beta\in C_\alpha} d_\delta\quad\forall\alpha\in\mathbf{K}.$$

Define

$$\Gamma\equiv\Gamma(\mu|\underset{d}{\Delta})=\sup_{\alpha\in\mathbf{K}}\Gamma_\alpha(\mu|\underset{d}{\Delta}); \tag{5.10}$$

then, for any fixed μ,

$$0<\Gamma<\infty. \tag{5.11}$$

It is not difficult to prove that $\Gamma(\mu)$ is a positive linear combination of $\cosh(k\mu)$, $k\in\mathbf{Z}^+$ by using the symmetry property (2.3a) and the definition of C_α. Therefore, $\Gamma(\mu)$ is even, monotonic increasing for $\mu>0$ and has a global minimum at $\mu=0$. It can also be proved that $\Gamma(0)=\sigma>0$, thus $\Gamma(\mu)$ can be represented as in Fig. 6. Examples 5.1 and 5.2 give the basic idea of the general discussion. Those results will be used later.

We now have

$$\begin{aligned}
&\frac{d}{dt}w_\alpha-\Lambda w_\alpha\\
&=A\frac{d}{dt}p(t)\,e^{-\mu(|\alpha|-ct)}+Ap(t)\,e^{-\mu(|\alpha|-ct)}\cdot\mu c-Ap(t)\,e^{-\mu(|\alpha|-ct)}\cdot\Gamma_\alpha\\
&=A\,e^{-\mu(|\alpha|-ct)}\left[\frac{d}{dt}p(t)+p(t)\mu c-\Gamma_\alpha p(t)\right]\\
&\geqq A\,e^{-\mu(|\alpha|-ct)}\left[\frac{d}{dt}p(t)+(\mu c-\Gamma)p(t)\right].
\end{aligned}$$

Let $p(t)$ be the solution of $(d/dt)p(t)+(\mu c-\Gamma)p(t)=0$; then

$$p(t)=e^{-\mu(c-\Gamma/\mu)t}. \tag{5.12}$$

So, by (5.6),

$$\frac{d}{dt}w_\alpha-\Lambda w_\alpha\geqq 0\geqq\frac{d}{dt}u_\alpha-\Lambda u_\alpha.$$

Furthermore, choose

$$A\equiv\sup_{\gamma\in\mathbf{Z}^d}\{e^{\mu|\gamma|}\cdot u_\gamma(0)\}. \tag{5.13}$$

A is finite because of the compact support assumption (5.2). Thus

$$\begin{aligned}
w_\alpha(0)&=Ap(0)\,e^{-\mu|\alpha|}\\
&=\sup_{\gamma\in\mathbf{Z}^d}\{e^{\mu|\gamma|}\cdot u_\gamma(0)\}\cdot 1\cdot e^{-\mu|\alpha|}\\
&\geqq u_\alpha(0).
\end{aligned}$$

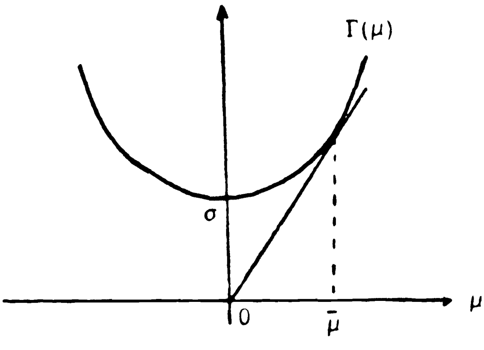


FIG. 6

This implies that $w_\alpha(t) \geqq u_\alpha(t)$ in R_o^+ for all $\alpha \in \mathbf{Z}^d$, because if choosing $f_\alpha \equiv 0$, $g_\alpha \equiv 0$, then all conditions of Lemma 2.1 are satisfied. From (5.7) and (5.12),

$$A\, e^{-\mu(|\alpha|-ct)} \cdot e^{-\mu(c-\Gamma/\mu)t} \geqq u_\alpha(t) \quad \text{in } R_o^+ \quad \text{for all } \alpha. \tag{5.14}$$

Now if μ and $\bar{c}$ are properly chosen, the left-hand side of (5.14) will go to zero if $c > \bar{c}$.

Let $\mu_o > 0$; the tangent line to $\Gamma(\mu)$, $\mu > 0$, at μ_o can be represented as

$$y(\mu) = \Gamma'(\mu_o)\mu + \Gamma(\mu_o) - \Gamma'(\mu_o)\mu_o .$$

Let $\bar{\mu} > 0$ be the value of μ_o such that the tangent line at $(\bar{\mu}, \Gamma(\bar{\mu}))$ passes through the origin (see Fig. 6). Then,

$$\Gamma(\bar{\mu}) = \Gamma'(\bar{\mu})\bar{\mu}. \tag{5.15}$$

The tangent line exists, because, as we discussed before, $\Gamma(\mu)$ is even, monotonic increasing for $\mu > 0$, and has a global minimum $\Gamma(0) = \sigma > 0$. Now choose

$$\bar{c} = \frac{\Gamma(\bar{\mu})}{\bar{\mu}} = \Gamma'(\bar{\mu}); \tag{5.16}$$

then $c > \Gamma(\bar{\mu})/\bar{\mu}$ if $c > \bar{c}$. But $\Gamma(\mu)/\mu$ is a continuous function of μ, so $c > \Gamma(\bar{\mu})/\bar{\mu}$ implies that there exists $\mu^+ > \bar{\mu}$ such that $c > \Gamma(\mu)/\mu$ for all μ with $\bar{\mu} \leqq \mu \leqq \mu^+$. Therefore, if $c > \bar{c}$ and μ is in the interval $(\bar{\mu}, \mu^+)$, then

$$\lim_{t\to\infty} e^{-\mu(c-\Gamma/\mu)t} = 0.$$

Furthermore, for any $\alpha_o \in \mathbf{Z}^d$, if $|\alpha| = |\alpha_o| + [ct]$, then the first factor in (5.14),

$$\begin{aligned} A\, e^{-\mu(|\alpha|-ct)} &= A\, e^{-\mu(|\alpha_o|+[ct]-ct)} \\ &\leqq A\, e^{-\mu(|\alpha_o|-1)}, \end{aligned}$$

is bounded.

Combining the results above, we obtain

$$\lim_{t\to\infty} A\, e^{-\mu(|\alpha|-ct)} \cdot e^{-\mu(c-\Gamma/\mu)t} = 0$$

if $|\alpha| = |\alpha_o| + [ct]$ for any $\alpha_o \in \mathbf{Z}^d$, $d > \bar{c}$ and $\mu \in (\bar{\mu}, \mu^+)$. The result of the theorem is obtained by considering the inequality given in (5.14) and $u_\alpha(t) \geqq 0$. □

Notice that there is no contradiction between our result and (5.1) because in (5.1) α is fixed, and in the result of Theorem 5.1, $\alpha = |\alpha_o| + [ct]$ is moving at a speed c.

Example 5.1. $d = 2$. Consider the Δ_d given in (2.7)

$$\underset{d}{\Delta} u_{i,j} = d_1(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + d_2(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}).$$

Here we have

$$\alpha = (i, j),$$

$$d_{\pm 1,0} = d_1 > 0, \quad d_{0,\pm 1} = d_2 > 0, \quad d_{\pm 1,\pm 1} = 0,$$

and from (5.9),

$$\begin{aligned}\Gamma_\alpha(\mu) &= d_1[e^{-\mu(|(i+1,j)|-|(i,j)|)} - k] + d_1[e^{-\mu(|(i-1,j)|-|(i,j)|)} - k]\\ &\quad + d_2[e^{-\mu(|(i,j+1)|-|(i,j)|)} - k] + d_2[e^{-\mu(|(i,j-1)|-|(i,j)|)} - k]\\ &= d_1[\underbrace{e^{-\mu(|(i+1,j)|-|(i,j)|)} + e^{-\mu(|(i-1,j)|-|(i,j)|)} - 2k}_{(1)}]\\ &\quad + d_2[\underbrace{e^{-\mu(|(i,j+1)|-|(i,j)|)} + e^{-\mu(|(i,j-1)|-|(i,j)|)} - 2k}_{(2)}],\end{aligned}$$

where

$$k = 1 - \frac{\sigma}{2(d_1 + d_2)} \quad \text{from (5.5).}$$

If $i = 0$, and $j = 0$, then

$$\begin{aligned}(1) &= e^{-\mu(1-0)} + e^{-\mu(1-0)} - 2k = 2(e^{-\mu} - k)\\ (2) &= 2(e^{-\mu} - k).\end{aligned}$$

If $i \leqq -1$, j is arbitrary, then

$$\begin{aligned}|(i+1,j)| - |(i,j)| &\equiv |i+1| + |j| - |i| - |j| = |i+1| - |i|\\ &= -(i+1) - (-i) = -1,\\ |(i-1,j)| - |(i,j)| &= |i-1| - |i| = -(i-1) - (-i) = 1,\end{aligned}$$

so

$$(1) = e^{-\mu} + e^{\mu} - 2k.$$

If $i \geqq 1$, j is arbitrary, then

$$\begin{aligned}|(i+1,j)| - |(i,j)| &= |i+1| - |i| = i + 1 - i = 1,\\ |(i-1,j)| - |(i,j)| &= |i-1| - |i| = i - 1 - i = -1,\end{aligned}$$

so

$$(1) = e^{\mu} + e^{-\mu} - 2k.$$

Similarly, if $j \leqq -1$ or $j \geqq 1$, and i is arbitrary, we have

$$(2) = e^{\mu} + e^{-\mu} - 2k.$$

Therefore, we obtain

$$\Gamma_{i,j} = \Gamma_{i,j}(\mu) \equiv \begin{cases} 2(d_1 + d_2)(e^{-\mu} - k) & \text{if } i = j = 0,\\ (d_1 + d_2)(e^{-\mu} + e^{\mu} - 2k) & \text{otherwise,}\end{cases}$$

and

$$\begin{aligned}\Gamma &= (d_1 + d_2)(e^{-\mu} + e^{\mu} - 2k)\\ &= (d_1 + d_2)\left(e^{-\mu} + e^{\mu} - 2 + \frac{\sigma}{d_1 + d_2}\right).\end{aligned} \tag{5.17}$$

Notice that $\min \Gamma = \sigma > 0$. Now (5.15) becomes

$$e^{-\bar{\mu}} + e^{\bar{\mu}} - 2 + \frac{\sigma}{d_1 + d_2} = (-e^{-\bar{\mu}} + e^{\bar{\mu}})\bar{\mu},$$

or

$$\cosh(\bar{\mu}) - \bar{\mu}\sinh(\bar{\mu}) + \frac{\sigma}{2(d_1 + d_2)} - 1 = 0. \tag{5.18}$$

This equation has a positive solution for any $d_1, d_2 > 0$. Finally, by (5.16) and (5.17),

$$\bar{c} = 2(d_1 + d_2)\sinh(\bar{\mu}). \tag{5.19}$$

□

Example 5.2. Use the operator Δ_d defined in (2.9), namely,

$$\begin{aligned}\underset{d}{\Delta} u_{i,j} &= d_1(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + d_2(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) \\ &\quad + d_3(u_{i+1,j+1} - 2u_{i,j} + u_{i-1,j-1}).\end{aligned}$$

Thus, from (5.9),

$$\begin{aligned}\Gamma_\alpha(\mu) = {} & \underbrace{d_1[e^{-\mu(|(i+1,j)|-|(i,j)|)} + e^{-\mu(|(i-1,j)|-|(i,j)|)} - 2k]}_{(1)} \\ & + \underbrace{d_2[e^{-\mu(|(i,j+1)|-|(i,j)|)} + e^{-\mu(|(i,j-1)|-|(i,j)|)} - 2k],}_{(2)} \\ & + \underbrace{d_3[e^{-\mu(|(i+1,j+1)|-|(i,j)|)} + e^{-\mu(|(i-1,j-1)|-|(i,j)|)} - 2k],}_{(3)}\end{aligned}$$

where

$$k = 1 - \frac{\sigma}{2(d_1 + d_2 + d_3)} \quad \text{from (5.5).}$$

As in Example 5.1, we have to go through various cases to determine (1), (2), and (3). From Example 5.1,

$$(1) = (2) = \begin{cases} 2(e^{-\mu} - k) & \text{if } i = j = 0, \\ e^{-\mu} + e^{\mu} - 2k & \text{otherwise.} \end{cases}$$

For (3) we have

$$(3) = \begin{cases} 2(e^{-2\mu} - k) & \text{if } i = j = 0, \\ e^{-2\mu} + e^{2\mu} - 2k & \text{if } i \geqq 1, j \geqq 1 \quad \text{or} \quad i \leqq -1, j \leqq -1, \\ 2(1 - k) & \text{if } i \geqq 1, j \leqq -1 \quad \text{or} \quad i \leqq -1, j \geqq 1. \end{cases}$$

Hence,

$$\Gamma_{i,j} = \begin{cases} 2(d_1 + d_2)(e^{-\mu} - k) + 2d_3(e^{-2\mu} - k) & i = j = 0, \\ (d_1 + d_2)(e^{-\mu} + e^{\mu} - 2k) + d_3(e^{-2\mu} + e^{2\mu} - 2k) & i \geqq 1, j \geqq 1 \quad \text{or} \\ & i \leqq -1, j \leqq -1, \\ (d_1 + d_2)(e^{-\mu} + e^{\mu} - 2k) + 2d_3(1 - k) & \text{otherwise.} \end{cases}$$

Since $\mu > 0$,

$$\begin{aligned}\Gamma(\mu) &= \sup_{(i,j) \in \mathbf{Z}^2} \Gamma_{i,j}(\mu) \\ &= (d_1 + d_2)(e^{-\mu} + e^{\mu} - 2k) + d_3(e^{-2\mu} + e^{2\mu} - 2k) \\ &= 2(d_1 + d_2)\cosh(\mu) + 2d_3\cosh(2\mu) - 2(d_1 + d_2 + d_3)k \\ &= 2[(d_1 + d_2)\cosh(\mu) + d_3\cosh(2\mu)] - 2(d_1 + d_2 + d_3) + \sigma,\end{aligned}$$

so that (5.15) becomes

$$\begin{aligned}2[(d_1+d_2)\cosh(\bar{\mu})+d_3\cosh(2\bar{\mu})]-2(d_1+d_2+d_3)+\sigma\\ =\bar{\mu}\{2(d_1+d_2)\cosh(\bar{\mu})+d_3\cosh(2\bar{\mu})\}.\end{aligned} \tag{5.20}$$

This equation has a positive solution $\bar{\mu}$. □

THEOREM 5.2. *Assume for* $\alpha\in\mathbf{Z}^d$,

(i) $\{q_\alpha\}$ *is not identical to* $\{0\}$ *and is a steady state solution to* (2.1)-(2.4) *with*

$$0\leqq q_\alpha<1 \quad \textit{for } |\alpha|<N<\infty, \quad q_\alpha\leqq 0 \quad \textit{for } |\alpha|=N;$$

(ii) *The only steady state* $\tau_\alpha\geqq 0$ for all $\alpha\in\mathbf{Z}^d$ *with* $\tau_\alpha\geqq q_\alpha$ *for* $|\alpha|<N$ *is* $\tau_\alpha=1$;

(iii) $\{u_\alpha(t)\}$ *is a solution to* (2.1)-(2.4), *and there exists* $\alpha_\circ\in\mathbf{Z}^d$, *such that*

$$u_{\alpha+\alpha_\circ}(0)\geqq q_\alpha \quad \textit{for } |\alpha|<N.$$

Then there exists $\underline{c}>0$, *such that for all* $c<\underline{c}$ *and* $\varepsilon>0$, *there exists* $T(\varepsilon,c)<\infty$, *with* $u_\alpha(t)\geqq 1-\varepsilon$ *for* $|\alpha|=|\alpha_\circ|+[ct]$ *and for* $t>T(\varepsilon,c)$.

Proof. We may assume that $\alpha_\circ=\mathbf{0}$, where $\mathbf{0}$ is the d-dimensional zero vector, and the general case can be obtained by a translation on α. The proof consists of the following.

(i) Let $\{w_\alpha(t)\}$ be the solution to (2.1) with $w_\alpha(0)=q_\alpha$ for $|\alpha|<N$, and $w_\alpha(0)=0$ for all α, $|\alpha|\geqq N$. By Lemma 4.4, $w_\alpha(t)\to 1$ monotonically for each α as $t\to\infty$. Thus, for each $\varepsilon>0$ there is a $t_1(\varepsilon)$, such that for $t\geqq t_1(\varepsilon)$,

$$w_{\mathbf{0}}(t)\geqq 1-\varepsilon. \tag{5.21}$$

(ii) Since there are only finitely many α's with $|\alpha|<N$, so that there exists $t_2<\infty$, such that for $t\geqq t_2$, and any α with $|\alpha|<N$, we have $w_{\alpha'}(t)\geqq w_\alpha(0)$, where $|\alpha'|=|\alpha|+1$. This is true because $w_\alpha(0)=q_\alpha<1$, and $w_{\alpha'}(t)\to 1$ as $t\to\infty$; therefore, for any α with $|\alpha|<N$ and any α' with $|\alpha'|=|\alpha|+1$, there exists $t'(\alpha,\alpha')$ such that $w_{\alpha'}(t)\geqq w_\alpha(0)$ for $t\geqq t'(\alpha,\alpha')$. Since there are only a finite number of $t'(\alpha,\alpha')$'s, we can choose

$$t_2=\max_{\alpha,\alpha'}\{t'(\alpha,\alpha')\}<\infty$$

$$\text{where } |\alpha|<N \quad\text{and}\quad |\alpha'|=|\alpha|+1.$$

Consider further that $w_\alpha(0)=0$ for $|\alpha|\geqq N$; we obtain

$$w_{\alpha'}(t)\geqq w_\alpha(0) \tag{5.22}$$

for all $\alpha\in\mathbf{Z}^d$, and all $\alpha'\in\mathbf{Z}^d$ with $|\alpha'|=|\alpha|+1$, as long as $t\geqq t_2$.

(iii) Equation (5.22) implies that

$$\begin{aligned}&w_\alpha(t+mt_2)\geqq w_{\mathbf{0}}(t)\\ &\text{for all } t>0 \quad\text{and}\quad m=|\alpha|\in\mathbf{Z}^+.\end{aligned} \tag{5.23}$$

In fact, from (5.22), $w_{\alpha'}(t_2)\geqq w_\alpha(0)$. By Lemma 2.2 (letting $\bar{w}_\alpha(0)\equiv w_{\alpha'}(t_2)$, etc.), we obtain

$$w_{\alpha'}(t+t_2)\geqq w_\alpha(t) \quad \text{for all } t\geqq 0. \tag{5.24}$$

Again, $w_{\alpha'}(2t_2)\geqq w_\alpha(t_2)$; so by Lemma 2.2,

$$w_{\alpha'}(t+2t_2)\geqq w_\alpha(t+t_2). \tag{5.25}$$

Let $\alpha=\mathbf{0}$; then from (5.24),

$$w_{\alpha'}(t+t_2)\geqq w_{\mathbf{0}}(t), \quad \text{where } |\alpha'|=|\mathbf{0}|+1=1. \tag{5.26}$$

In (5.25), letting $|\alpha|=1$, we obtain

$$w_{\alpha'}(t+2t_2) \geqq w_\alpha(t+t_2), \quad \text{where } |\alpha'|=|\alpha|+1=2,$$
$$\geqq w_0(t).$$

So (5.23) holds for $m=2$. Equation (5.23) can be obtained by using induction on m. Write (5.23) as

$$w_\alpha(t) \geqq w_0(t-mt_2) \quad \text{for } t \geqq mt_{2'} \quad \text{where } |\alpha|=m. \tag{5.27}$$

(iv) Let $\underline{c}=1/t_2$, and consider $w_\alpha(t)$ where $|\alpha|=[ct]$. If $c<\underline{c}$ then $1-ct_2=1-(c/\underline{c})>0$. Let $m=[ct]$ and $T(\varepsilon, c)=t_1(\varepsilon)/(1-ct_2)>0$ in (5.27); then for $t>T(\varepsilon, c)$,

$$w_\alpha(t) \geqq w_0(t-[ct]t_2) \geqq w_0(t(1-ct_2)) \geqq w_0(T(1-ct_2)) = w_0(t_1(\varepsilon))$$
$$\geqq 1-\varepsilon, \quad \text{where } |\alpha|=m=[ct].$$

Finally, the assumption $u_\alpha(0) \geqq q_\alpha = w_\alpha(0)$ implies, by Lemma 2.2, that $u_\alpha(t) \geqq w_\alpha(t)$ for all $t \geqq 0$, so $u_\alpha(t) \geqq 1-\varepsilon$ for $t>T(\varepsilon, c)$ and $c<\underline{c}$. □

Example 5.3. Returning to Example 5.1, let us calculate $\bar{c}$ and $\underline{c}$ for the system

$$\frac{d}{dt}u_{i,j} = d_1(u_{i+1,j}-2u_{i,j}+u_{i-1,j}) + d_2(u_{i,j+1}-2u_{i,j}+u_{i,j-1}) + f_{i,j}(u_{i,j}).$$

In order to simplify the calculations, let $f_{i,j}(u)=f(u)$, for all i, j, and let

$$f(u)=\begin{cases} -\frac{65}{62}u & u<\frac{62}{650} \\ \frac{65}{3}(u-0.1) & \frac{62}{650} \leqq u < \frac{71}{260} \\ 3.75 & \frac{71}{260} \leqq u < 0.75 \\ 15(1-u) & 0.75 \leqq u. \end{cases} \tag{5.28}$$

$f(u)$ is shown in Fig. 7. It can be treated as an approximation of some C^1 function.

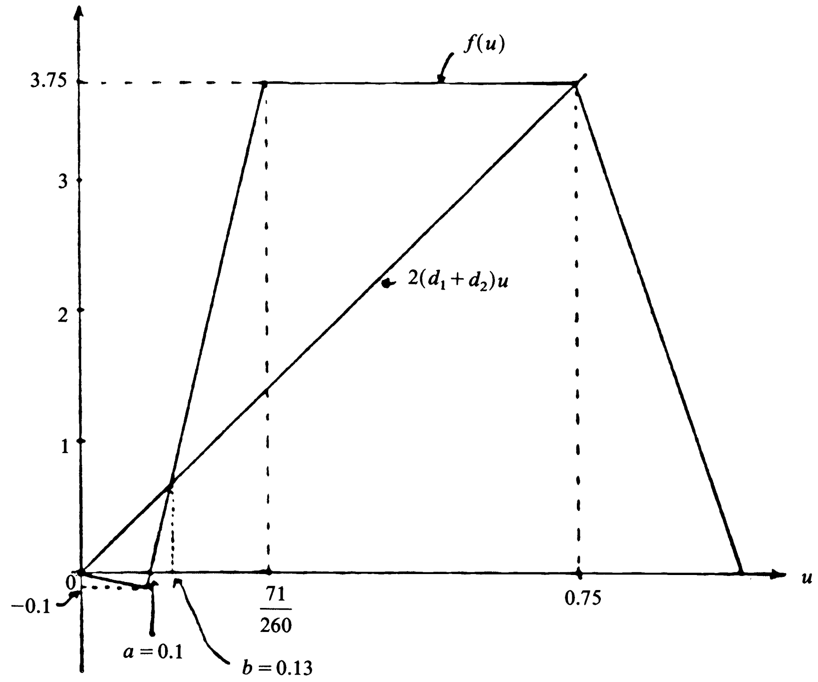


FIG. 7

Letting $d_1=1$, $d_2=1.5$, then (4.4), (4.5), and (4.6) are satisfied with $m=0.1$, $M=3.75$, $a=0.1$, $b=0.13$, and $e=0.75$. If we choose

$$u_{i,j}(0)=\begin{cases}0 & \text{if } i\neq 0,\ j\neq 0,\\ 0.13 & \text{if } i=j=0,\end{cases} \tag{5.29}$$

then by Theorem 4.5, $\lim_{t\to\infty} u_{i,j}(t)=1$.

The solution has been calculated numerically. We now calculate the upper bound and lower bound of the propagation speed.

(i) Calculation of the upper bound $\bar{c}$.

By (5.19), where $\bar{\mu}$ can be solved from (5.18),

$$\cosh(\bar{\mu})-\bar{\mu}\cdot\sinh(\bar{\mu})+\frac{\sigma}{2(d_1+d_2)}-1=0.$$

Now

$$\sigma_1=\sup_{0\leqq u<62/650}\left\{\frac{f(u)}{u}\right\}=\sup_{0\leqq u<62/650}\left\{-\left(\frac{\frac{65}{62}u}{u}\right)\right\}=-\frac{65}{62}.$$

Similarly, $\sigma_2=\sigma_3=975/71$, $\sigma_4=5$, so that

$$\sigma=\sup_{0\leqq u\leqq 1}\left\{\frac{f(u)}{u}\right\}$$
$$=\max\{\sigma_1,\sigma_2,\sigma_3,\sigma_4\}=\frac{975}{71}.$$

Hence $\cosh(\bar{\mu})-\bar{\mu}\cdot\sinh(\bar{\mu})+124/71=0$, which has the positive solution $\bar{\mu}\cong 1.715867$. Therefore, using values of d_j from Example 5.3, we have

$$\bar{c}=2(d_1+d_2)\sinh(\bar{\mu})\cong 5\sinh(1.715867)$$
$$\cong 13.454219.$$

(ii) Calculation of the lower bound $\underline{c}$.

$\underline{c}$ can be estimated by using the method used in the proof of Theorem 5.2.

We can choose $\underline{c}=1/t_2$, where $t_2=\max_{i,j,i'j'}\{t'(i,j,i',j')\}$ can be estimated as follows. From the given initial condition (5.28), we can see that $N=1$, $q_{0,0}=0.13$, and $q_{0,\pm 1}=q_{\pm 1,0}=0$, so that $w_{0,0}=0.13$, and $w_{i,j}=0$ if $i\neq 0$ and $j\neq 0$. From (2.1), letting $(i,j)=(0,1)$ and $t=0$, we obtain

$$\frac{d}{dt}w_{0,1}|_{t=0}=d_1(w_{1,1}-2w_{0,1}+w_{-1,1})+d_2(w_{0,2}-2w_{0,1}+w_{0,0})+f(w_{0,1})|_{t=0}$$
$$=d_2w_{0,0}(0)=1.5\times 0.13$$
$$=0.195.$$

If we assume that $w_{0,1}(t)$ is increasing linearly for small t (we do not know how small t must be, so this is only a rough approximation), then it will take $0.13/0.195\cong 0.667$ unit time to grow from 0 to 0.13. Similarly,

$$\frac{d}{dt}w_{1,0}|_{t=0}=d_1w_{0,0}(0)=0.13,$$

thus it will take 1 unit time for $w_{1,0}$ to grow from 0 to 0.13. Using the same calculation on $w_{-1,0}$, and $w_{0,-1}$, we obtain similar results. Therefore, $t_2 = \max(0.667, 1) = 1$, thus $\underline{c} = 1/t_2 = 1$.

6. Numerical simulations. Let us consider the system discussed in Examples 5.1 and 5.3, i.e.,

$$\frac{d}{dt} u_{i,j} = d_1(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + d_2(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) + f(u_{i,j})$$

$$(i,j) \in \mathbf{Z}^2,$$

where $f(u)$ is given in (5.28). Let $d_1 = 1$, $d_2 = 1.5$; then (4.4), (4.5), and (4.6) are satisfied. We also set $u_{i,j}(t) \equiv 0$ on the boundary of the grid.

(1) Choosing initial condition

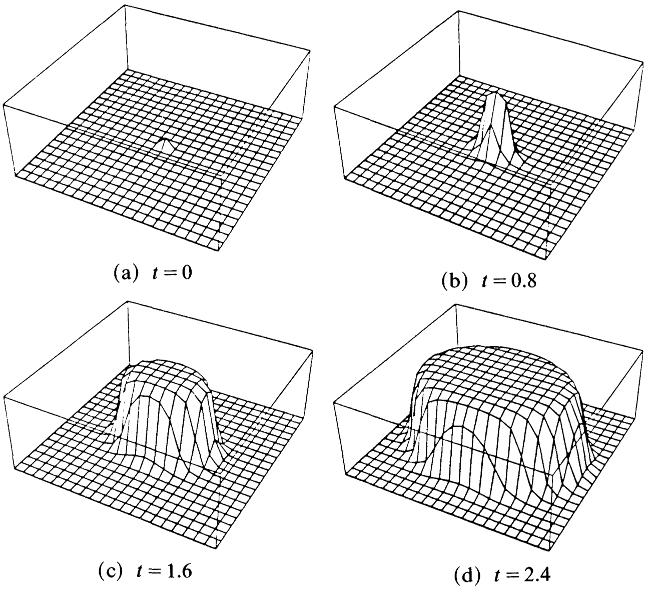
$$u_{i,j}(0) = 0.095 \quad \text{for all } i,j \text{ (not on the boundary).}$$

$u_{i,j}(0) < a$ for all i,j. From the numerical simulations, it is clear that $u_{i,j}(t) \to 0$ for all i,j when $t \to \infty$.

(2) Choosing

$$u_{i,j} = \begin{cases} 0 & \text{if } i \neq 0, j \neq 0, \\ 0.13 & \text{if } i = j = 0. \end{cases}$$

The result is shown in Fig. 8. From these graphs, we can see that $u_{i,j}(t) \to 1$ as $t \to \infty$, for all $i,j \in$ interior and $\notin$ boundary of the grid, which supports our superthreshold result, Theorem 4.5.



(a) $t = 0$

(b) $t = 0.8$

(c) $t = 1.6$

(d) $t = 2.4$

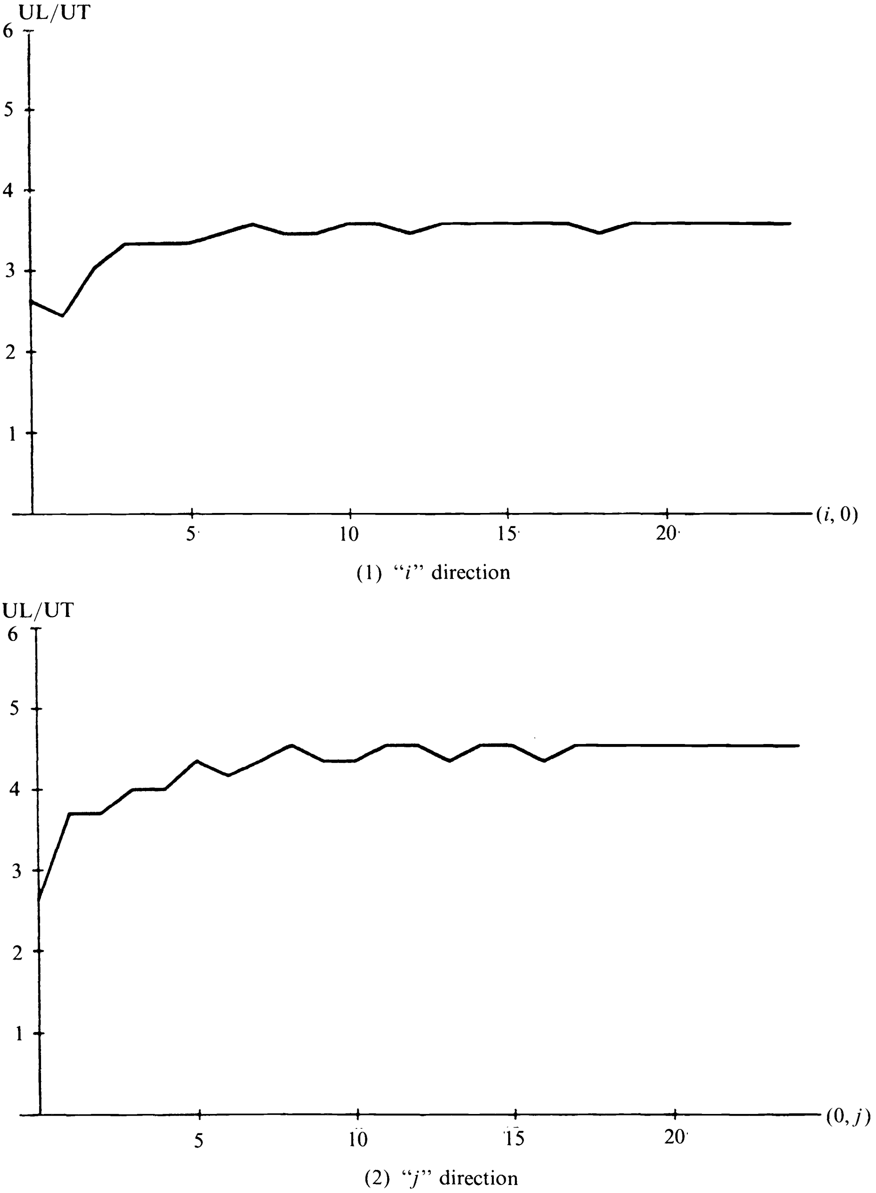
FIG. 8. *Superthreshold experiment for 2-D initial value problem.*

$$u_{i,j}(0) = \begin{cases} 0 & (i,j) \neq (0,0) \\ 0.13 & (i,j) = (0,0) \end{cases}, \qquad t_f = 2.4.$$

Propagation speed in two particular directions, the "i" direction (set $j \equiv 0$) and the "j" direction (set $i \equiv 0$), are shown in Fig. 9. From Fig. 9(1), along the "i" direction, after the node (5, 0), the propagation speed starts to stabilize at about 3.5 UL/UT (Unit Length/Unit Time). Along the "j" direction, from Fig. 9(2), the propagation speed starts to stabilize at about 4.5 UL/UT after the node (0, 7). In Example 2.3, we obtained $\underline{c} \approx 1$, $\bar{c} \approx 13.5$; therefore,

$$\underline{c} < c \text{ (both directions)} < \bar{c}.$$

The theoretical values of $\bar{c}$ and $\underline{c}$ are reasonable compared with the experimental result. It is seen that the comparison technique used gives rather conservative bounds. Here



(1) "i" direction

(2) "j" direction

FIG. 9. *Propagation speed.*

TABLE 1
Propagation tests.

$\mathbf{\Omega}$	$u_0(t)$	σ	Propagation
$\mathbf{\Omega}_1$	0.13	1.0	pass
	0.13	10.0	pass
$\mathbf{\Omega}_2$	0.13	0.05	not pass
	0.3	0.1	not pass
$\mathbf{\Omega}_3$	0.13	1.0	pass
	0.13	10.0	not pass

we choose $d_1 = 1.0$ ("i" direction), $d_2 = 1.5$ ("j" direction), the above calculation suggests that the larger d is, the higher the propagation speed in that corresponding direction.

We also considered propagation behavior of the inhomogeneous system

$$\frac{d}{dt} u_{i,j} = d_1(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + d_2(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) + f_{i,j}(u_{i,j}),$$

with

$$f_{i,j}(u) = \begin{cases} f(u) & (i,j) \notin \mathbf{\Omega}, \\ -\sigma u & (i,j) \in \mathbf{\Omega}, \end{cases}$$

where $\mathbf{\Omega} \subset \mathbf{Z}^2$; $f(u)$ is given in (5.28) which satisfies the "propagation conditions."

The above function $f_{i,j}$ is motivated by considering the thin heart tissue model with the cells in region $\mathbf{\Omega}$ damaged. We want to study when propagation can pass through the damaged region. We choose $d_1 = 1.0$, $d_2 = 1.5$, grid size 70 × 70, and terminal time $t_f = 8.0$. The effects of initial conditions, σ, and shape of the damaged region on propagation was considered. The result is summarized as the following:

$$u_{i,j}(0) = \begin{cases} 0 & \text{if } i \neq 0, j \neq 0, \\ u_0 & \text{if } i = j = 0, \end{cases}$$

$$\mathbf{\Omega}_1 = \{(i,j) | 3 \leqq i \leqq 9, -3 \leqq j \leqq 3\} \qquad (3 \times 3 \text{ square centered at } (6,0)),$$

$$\mathbf{\Omega}_2 = \{(i,j) | 4 \leqq j \leqq 8\} \qquad (\text{a vertical band with width } 4),$$

$$\mathbf{\Omega}_3 = \{(i,j) | \{|i| \leqq 10 \cap |j| \leqq 10\} \backslash \{|i| \leqq 6 \cap |j| \leqq 6\}\}$$

(a rectangular ring centered at (0, 0) with center radius 8 and width 4).

The calculations were done using a fourth-order Runge-Kutta method on a VAX computer. Graphs were plotted using the package "Mathematical." Step size was a uniform .01 time units with grid size 70 × 70. For some calculation, in order to obtain a stable pattern, we need to increase the terminal time t_f; in such a case, to compensate the usage of CPU time, the grid size is reduced.

Acknowledgment. The author is grateful to Professor Jonathan Bell for many stimulating discussions and useful advice as well as the help in revising the manuscript.

REFERENCES

[1] D. ARONSON AND H. WEINBERGER, *Nonlinear diffusion in population genetics, combustion, and nerve pulse propagation*, in Lecture Notes in Math. 446, Springer-Verlag, Berlin, 1975, pp. 5-49.

[2] J. BELL AND C. COSNER, *Threshold behavior and propagation for nonlinear differential-difference systems motivated by modeling myelinated axons*, Quart. App. Math., 42 (1984), pp. 1-14.

[3] J. BELL, *Some threshold results for models of myelinated nerves*, Math. Biosci. 54 (1981), pp. 181-190.

[4] ———, *Behavior of some models of myelinated axons*, IMA J. Math. Appl. Med. Biol. 1 (1984), pp. 149-167.

[5] W. GAO, *Long time behavior of a class of semidiscrete dynamic systems*, Ph.D. dissertation, State University of New York at Buffalo, Buffalo, NY, 1990.

[6] J. KEENER, *Propagation and its failure in coupled systems of discrete excitable cells*, SIAM J. Appl. Math, 47 (1987), pp. 556-572.

[7] ———, *A mathematical model for the vulnerable phase in myocardium*, Math. Biosci., 90 (1988), pp. 3-18.

[8] B. ZINNER, *Existence of traveling wavefront solutions for a discrete Nagumo equation*, preprint.

[9] ———, *Stability of traveling wavefronts for the discrete Nagumo equation*, preprint.

GLOBAL ASYMPTOTIC STABILITY FOR A STATIONARY SOLUTION OF A SYSTEM OF INTEGRO-DIFFERENTIAL EQUATIONS DESCRIBING THE FORMATION OF LIVER ZONES*

KJELL HOLMÅKER†

Abstract. The formation of liver zones is modeled by a system of integro-differential equations. It has previously been proved that one particular stationary solution, characterized by a jump discontinuity at the zone boundary, is asymptotically stable with respect to sufficiently small perturbations of a special type. In this paper it is proved that this stationary solution is in fact globally asymptotically stable, that is, it is the limit (as time tends to infinity) of the solution of the integro-differential equations for arbitrary initial values.

Key words. integro-differential equations, global asymptotic stability, self-organization of cellular patterns

AMS(MOS) subject classifications. 45K05, 45M05, 92C15

1. Introduction.

1.1. Biological background.

The liver performs its metabolic functions with the aid of various enzymes fixed inside liver cells. These immobile cells line the many capillaries (hepatic sinusoids) through which the total hepatic blood flow is manifolded, whereby exchange of substances between blood and the cells is facilitated. The interplay of the unidirectional blood flow with local metabolism generates concentration gradients of blood-borne substances (such as oxygen) between the inlet and the outlet of the liver.

Several metabolic functions of the liver have been found to be organized in spatial zones arranged in relation to the direction of hepatic blood flow, in such a way that some enzymes act almost wholly upstream of others. We shall attribute such distributions of enzyme activities to distributions of cell-types. For the simplest case of two enzymes, let there be two corresponding cell-types, each containing only one of the enzymes; separate metabolic zones occur when all cells of the one type are located upstream of all cells of the other type. We shall suppose, furthermore, that each cell-type reproduces itself by division.

It has been shown recently that, for several kinds of enzyme kinetics, the observed zonal structure would result from the implementation of a certain physiologically desirable optimization principle [2], [3], [4], [5]. A mechanism by which that structure could develop in any one liver was considered in [1]. It is formulated in terms of competitive exclusion of the two cell-types in space and time and is an example of self-organization of gross cellular patterns. The mathematical model was discussed in [1], but for convenience the main steps in its derivation are repeated here.

As the many capillaries comprising the liver are similar and act essentially in parallel, we shall model a representative capillary lined with cells of two kinds. We put the x-axis along the blood flow, with inlet at $x=0$ and outlet at $x=\ell$. We define the density of cells of the first kind, $\rho_1(t,x)$, as a continuous representation of the number of cells of the first kind per unit length of capillary at time t at the position x. The density $\rho_2(t,x)$ of cells of the second kind is defined analogously. The total cell density $\rho_1+\rho_2$ cannot exceed some fixed maximum density σ of cell sites, as division of the cells is limited by the familiar phenomenon of contact inhibition.

* Received by the editors March 26, 1990; accepted for publication (in revised form) June 8, 1992.

† Department of Mathematics, Chalmers University of Technology, and the University of Göteborg, S-412 96 Göteborg, Sweden.

The local rate of change $\partial\rho_1/\partial t$ of the density of cells of the first kind is assumed to consist of a growth-rate term proportional to ρ_1 (self-generation) and to the density of sites available, $\sigma-\rho_1-\rho_2$; and of a death-rate term proportional to ρ_1, with a coefficient $\beta_1(c)>0$ dependent on the local concentration c of a controlling blood-borne substance (we consider oxygen). We shall assume that $\beta_1(c)$ is of the form

$$\beta_1(c)=\mu_1-\nu_1(c-c_0),$$

where c_0 is the steady oxygen concentration at the inlet, and $\mu_1\geqq 0$, $\nu_1\geqq 0$ are constants. Then

$$\frac{\partial\rho_1}{\partial t}=k_1\rho_1(\sigma-\rho_1-\rho_2)-(\mu_1-\nu_1(c-c_0))\rho_1,$$

where $k_1>0$ is a constant. A similar equation is obtained for ρ_2. Oxygen is transported in the x-direction predominantly by convection with the blood and is used up by the two cell-types at the rates $\kappa_1\rho_1$ and $\kappa_2\rho_2$ (with positive constants κ_1 and κ_2). Changes in c caused by changes in ρ_1 and ρ_2 are quasi-steady. Therefore, c satisfies

$$f\frac{\partial c}{\partial x}=-\kappa_1\rho_1-\kappa_2\rho_2,$$

where f is the steady rate of blood flow. If this is integrated and c eliminated, we obtain the equations

$$\frac{\partial\rho_1}{\partial t}=\rho_1\left[k_1(\sigma-\rho_1-\rho_2)-\mu_1-\frac{\nu_1}{f}\int_0^x(\kappa_1\rho_1+\kappa_2\rho_2)\,d\xi\right],$$
$$\frac{\partial\rho_2}{\partial t}=\rho_2\left[k_2(\sigma-\rho_1-\rho_2)-\mu_2-\frac{\nu_2}{f}\int_0^x(\kappa_1\rho_1+\kappa_2\rho_2)\,d\xi\right].$$

If $k_1\sigma\leqq\mu_1$, then $\rho_1(t,x)\to 0$ as $t\to\infty$ for all x, and similarly for ρ_2. Therefore, we assume that $k_1\sigma>\mu_1$ and $k_2\sigma>\mu_2$. For similar reasons we also assume that at least one of ν_1 and ν_2 is positive (say ν_1).

To simplify the equations we introduce new variables

$$t'=(k_1\sigma-\mu_1)t,\quad x'=\frac{\nu_1\kappa_1}{fk_1}x,\quad v_i(t',x')=\frac{k_1}{c_1}\rho_i(t,x),\quad i=1,2,$$

and new parameters

$$\theta=\frac{\kappa_2}{\kappa_1},\quad \gamma=\frac{k_2}{k_1},\quad \lambda=\frac{k_1(k_2\sigma-\mu_2)}{k_2(k_1\sigma-\mu_1)},\quad \eta=\frac{\nu_2k_1}{\nu_1k_2}.$$

After dropping the primes we obtain

$$\begin{aligned}\frac{\partial v_1}{\partial t}(t,x)&=v_1(t,x)\left[1-v_1(t,x)-v_2(t,x)-\int_0^x[v_1(t,\xi)+\theta v_2(t,\xi)]\,d\xi\right],\\ \frac{\partial v_2}{\partial t}(t,x)&=\gamma v_2(t,x)\left[\lambda-v_1(t,x)-v_2(t,x)-\eta\int_0^x[v_1(t,\xi)+\theta v_2(t,\xi)]\,d\xi\right].\end{aligned}\tag{1.1}$$

This system is studied for $t\geqq 0$, $0\leqq x\leqq L$, where $L=\nu_1\kappa_1\ell/fk_1$, with given initial values

$$v_i(0,x)=v_i^0(x),\quad i=1,2,\quad 0\leqq x\leqq L.\tag{1.2}$$

The parameters satisfy

$$\theta>0,\quad \gamma>0,\quad \lambda>0,\quad \eta\geqq 0,\tag{1.3}$$

and the initial functions v_i^0 are nonnegative, bounded, and measurable functions.

1.2. The mathematical problem. In [1] it was shown that (1.1), (1.2) has a unique solution (v_1, v_2) such that the functions $v_i = v_i(t, x)$ are continuously differentiable in t for each x, measurable in x for each t, and satisfy (1.1), (1.2) for each (t, x). Furthermore, there is a constant $C \geqq 1$ such that

$$0 \leqq v_i(t, x) \leqq C \quad \text{for all } t \geqq 0, \quad x \in [0, L], \quad i = 1, 2. \tag{1.4}$$

The main question is what happens to the solution (v_1, v_2) as $t \to \infty$. One may guess that it tends to a stationary solution of (1.1), that is to a nonnegative solution of the system

$$\left.\begin{aligned} v_1\left[1 - v_1 - v_2 - \int_0^x (v_1 + \theta v_2)\, d\xi\right] &= 0, \\ v_2\left[\lambda - v_1 - v_2 - \eta \int_0^x (v_1 + \theta v_2)\, d\xi\right] &= 0, \end{aligned}\right\} \quad 0 \leqq x \leqq L.$$

It was shown in [1] that all stationary solutions are unstable except possibly one (see [1] for the definition of stability for this problem). In the most interesting case, where the parameters satisfy

$$\eta < \lambda < 1, \qquad 0 < x^* < L, \tag{1.5}$$

where

$$x^* = \ln \frac{1 - \eta}{\lambda - \eta}, \tag{1.6}$$

this stationary solution is

$$v_1^s(x) = \begin{cases} e^{-x} & \text{for } 0 \leqq x < x^*, \\ 0 & \text{for } x^* < x \leqq L, \end{cases} \tag{1.7}$$

$$v_2^s(x) = \begin{cases} 0 & \text{for } 0 \leqq x < x^*, \\ e^{-x^*} e^{-\eta\theta(x - x^*)} & \text{for } x^* < x \leqq L, \end{cases} \tag{1.8}$$

a solution that has the desired zonal structure. It was also shown in [1] that the stationary solution (v_1^s, v_2^s) is asymptotically stable with respect to a certain class of perturbations. Apart from being sufficiently small (uniformly in x) these perturbations had to go to zero at a certain rate as $x \to x^*$. In [1] the question of what happens for more general perturbations was left open, let alone the question of how the solution of (1.1), (1.2) behaves for initial functions (1.2) that are not necessarily close to the equilibrium solution (1.7), (1.8). We can now answer these questions in that we can show that $v_i(t, x) \to v_i^s(x)$ as $t \to \infty$ for $x \neq x^*$, $i = 1, 2$, for arbitrary initial functions v_i^0 (bounded away from zero). In other words we have the following theorem.

THEOREM 1. *Assume that the parameters satisfy* (1.3) *and* (1.5). *Let v_i^0 be bounded, measurable functions on* $[0, L]$ *such that*

$$v_i^0(x) \geqq \alpha_0 > 0 \quad \textit{for all } x \in [0, L], \quad i = 1, 2, \tag{1.9}$$

for some constant α_0. Then

$$v_i(t, x) \to v_i^s(x) \quad \textit{as } t \to \infty \quad \textit{for } x \neq x^*, \quad i = 1, 2, \tag{1.10}$$

where (v_1, v_2) is the solution of (1.1), (1.2), *and (v_1^s, v_2^s) is defined by* (1.7), (1.8). *The convergence in* (1.10) *is exponential for both v_1 and v_2 if $x < x^*$, and for v_1 if $x > x^*$.*

2. Proof of Theorem 1. The proof, which is rather long, is divided into three parts that are given as separate lemmas. In Lemma 2.1 we prove convergence for v_1 and v_2 if $x<x^*$, in Lemma 2.2 for v_1 if $x>x^*$, and in Lemma 2.3 for v_2 if $x>x^*$. In each lemma the proof is preceded by a comment where the main ideas of the proof are outlined.

LEMMA 2.1. *Let a, $0<a<x^*$, be fixed. Then there exist positive constants β and ω (depending on a) such that*

$$|v_1(t,x)-e^{-x}|\leqq \beta e^{-\omega t},\qquad v_2(t,x)\leqq \beta e^{-\omega t} \tag{2.1}$$

for $t\geqq 0$ and $x\in[0,a]$. If a is so close to x^* that $\nu(a)=\gamma[(1-\eta)e^{-a}+\eta-\lambda]<e^{-a}$, then ω can actually be taken as $\nu(a)$.

Comment. First, (2.1) can be obtained if x is sufficiently small. Using this result we can then extend (2.1) to a somewhat larger interval. Continuing in this fashion, it is possible to proceed step by step along the x-axis. Because of the positive difference x^*-a, we can reach the point a after a finite number of steps (the number depending on a). In each step the estimate for v_2 is obtained first. As for v_1, we must first prove that it remains bounded away from zero on the new interval. Then comes the crucial step of the proof, which is the introduction of a Lyapunov type function. From a differential inequality for this function a convergence result for v_1 in L_2-norm is obtained. Finally, another differential inequality, derived from the differential equation for v_1, gives the pointwise estimate in (2.1).

Proof of Lemma 2.1. Since $1-\lambda-(1-\eta)\int_0^{x^*}e^{-\xi}\,d\xi=0$, we have

$$\delta_1=1-\lambda-(1-\eta)\int_0^a e^{-\xi}\,d\xi>0. \tag{2.2}$$

Choose a $\delta>0$ such that

$$\delta\leqq\min\left[\frac{\delta_1}{2(1-\eta)(1+\theta)(x^*+C)},\frac{\lambda-\eta}{2(1-\eta)(1+(1+\theta)x^*+C)}\right], \tag{2.3}$$

where C is the constant from (1.4), and an integer N such that $N\delta=a$. Divide the interval $[0,a]$ onto N parts $[x_{k-1},x_k]$, $k=1,\dots,N$, where $x_k=k\delta$, and let

$$\omega'=\frac{\gamma\delta_1}{2}. \tag{2.4}$$

Suppose that for a certain $k-1$, $1\leqq k-1\leqq N-1$, we have found positive numbers β_{k-1} and ω_{k-1} such that

$$|v_1(t,x)-e^{-x}|\leqq\beta_{k-1}e^{-\omega_{k-1}t}, \tag{2.5}$$

$$v_2(t,x)\leqq\beta_{k-1}e^{-\omega' t} \tag{2.6}$$

for $t\geqq 0$, $x\in[0,x_{k-1}]$. We are going to show that for some $\beta_k\geqq\beta_{k-1}$ and $0<\omega_k\leqq\omega_{k-1}$ we have

$$|v_1(t,x)-e^{-x}|\leqq\beta_k e^{-\omega_k t}, \tag{2.7}$$

$$v_2(t,x)\leqq\beta_k e^{-\omega' t} \tag{2.8}$$

for $t\geqq 0$, $x\in[0,x_k]$. This will also be obtained for $k=1$ without any extra assumptions.

From (2.5) and (2.6) we see that there is a T_k such that

$$|v_1(t,x)-e^{-x}|\leqq\delta,\ v_2(t,x)\leqq\delta\quad\text{for } t\geqq T_k,\ x\in[0,x_{k-1}]. \tag{2.9}$$

On $(x_{k-1}, x_k]$ we get for $t \geqq T_k$ (if $k=1$ we consider $[0, x_1]$ and $t \geqq 0$),

(2.10)
$$\begin{aligned}
&1-\lambda-(1-\eta)\int_0^x (v_1+\theta v_2)\,d\xi \\
&\quad\geqq 1-\lambda-(1-\eta)\int_0^{x_{k-1}}(v_1+\theta v_2)\,d\xi-(1-\eta)\int_{x_{k-1}}^{x_k}(v_1+\theta v_2)\,d\xi \\
&\quad\geqq 1-\lambda-(1-\eta)\int_0^{x_{k-1}}(v_1-e^{-\xi}+\theta v_2)\,d\xi-(1-\eta)\int_0^{x_{k-1}}e^{-\xi}\,d\xi-(1-\eta)(1+\theta)C\delta \\
&\quad\geqq 1-\lambda-(1-\eta)(1+\theta)\delta x^*-(1-\eta)\int_0^a e^{-\xi}\,d\xi-(1-\eta)(1+\theta)C\delta \\
&\quad=\delta_1-(1-\eta)(1+\theta)(x^*+C)\delta \geqq \delta_1-\frac{\delta_1}{2}=\frac{\delta_1}{2},
\end{aligned}$$

where we have used (1.4), (2.9), (2.2), and (2.3). From (1.1), (1.2), and (1.9) it follows that $v_i > 0$, $i = 1, 2$, and

$$\frac{\partial}{\partial t}\ln\frac{v_2}{v_1^\gamma}=\gamma\left[\lambda-1+(1-\eta)\int_0^x (v_1+\theta v_2)\,d\xi\right],$$

so that

(2.11)
$$\frac{v_2(t,x)}{[v_1(t,x)]^\gamma}=\frac{v_2^0(x)}{[v_1^0(x)]^\gamma}\,e^{-\gamma\int_0^t[1-\lambda-(1-\eta)\int_0^x(v_1(\tau,\xi)+\theta v_2(\tau,\xi))\,d\xi]\,d\tau}.$$

From (1.4), (1.9), (2.10), (2.11), and (2.4) we see that

$$v_2(t,x) \leqq C\left(\frac{C}{\alpha_0}\right)^\gamma e^{\gamma[\lambda-1+(1-\eta)(1+\theta)Cx^*]T_k}\,e^{-\omega'(t-T_k)}$$

for $t \geqq T_k$, $x \in (x_{k-1}, x_k]$. But for $0 \leqq t \leqq T_k$ we have $v_2(t,x) \leqq C\,e^{\omega' T_k}\,e^{-\omega' t}$, so that there is a β'_k such that

(2.12)
$$v_2(t,x) \leqq \beta'_k\,e^{-\omega' t} \quad \text{for } t \geqq 0, \quad x \in (x_{k-1}, x_k].$$

Then choose $T'_k \geqq T_k$ so that $\beta'_k\,e^{-\omega' T'_k} \leqq \delta$. Thus

(2.13)
$$v_2(t,x) \leqq \delta \quad \text{for } t \geqq T'_k, \quad x \in [0, x_k].$$

As a first step towards proving (2.7) we want to show that $v_1(t,x) \geqq \alpha_k$ for $t \geqq T'_k$, $x \in (x_{k-1}, x_k]$, where

(2.14)
$$\alpha_k=\min\left[\frac{\lambda-\eta}{2(1-\eta)},\ \alpha_0\,e^{(1-2C-(1+\theta)Cx^*)T'_k}\right].$$

Assume on the contrary that $v_1(\bar t, \bar x) < \alpha_k$ for some $\bar t \geqq T'_k$ and $\bar x \in (x_{k-1}, x_k]$. As long as $v_1(t, \bar x) < \alpha_k$ (and $t \geqq T'_k$), we have, from (2.9), (2.13), (2.14), (1.6), and (2.3),

$$\begin{aligned}
&1-v_1-v_2-\int_0^{\bar x}(v_1+\theta v_2)\,d\xi \\
&\quad> 1-\alpha_k-\delta-\int_0^{x_{k-1}}(v_1-e^{-\xi})\,d\xi-\int_0^{x_{k-1}}e^{-\xi}\,d\xi-\int_{x_{k-1}}^{x_k}v_1\,d\xi-\int_0^{x_k}\theta v_2\,d\xi \\
&\quad> 1-\alpha_k-\delta-\delta x^*-\int_0^{x^*}e^{-\xi}\,d\xi-C\delta-\theta\delta x^* \\
&\quad=\frac{\lambda-\eta}{1-\eta}-(1+(1+\theta)x^*+C)\delta-\alpha_k \geqq \frac{\lambda-\eta}{1-\eta}-\frac{\lambda-\eta}{2(1-\eta)}-\alpha_k \geqq 0.
\end{aligned}$$

Therefore, $(\partial v_1/\partial t)(t, \bar{x}) > 0$ for these t. We have, from (1.1), (1.9), and (2.14),

$$v_1(t, x) \geqq \alpha_0 \, e^{[1-2C-(1+\theta)Cx^*]T_k'} \geqq \alpha_k$$

for $t \leqq T_k'$ and $x \in [0, x^*)$. In particular, $v_1(T_k', \bar{x}) \geqq \alpha_k$, and, therefore, there is an interval $[t', t'']$ with $t' \geqq T_k'$ such that $v_1(t', \bar{x}) = \alpha_k$, and $v_1(t, \bar{x}) < \alpha_k$ for $t \in (t', t'']$. But this is a contradiction, since $(\partial v_1/\partial t)(t, \bar{x}) > 0$ on $(t', t'']$. Thus $v_1(t, x) \geqq \alpha_k$ for all $t \geqq 0, x \in (x_{k-1}, x_k]$.

Let

$$V_k(t) = \int_{x_{k-1}}^{x_k} [v_1(t, x) - e^{-x} - e^{-x} \ln v_1(t, x) - x\, e^{-x}] \, dx,$$

$$w_1(t, x) = v_1(t, x) - e^{-x},$$

and

$$\psi(t, x) = \int_0^x [v_1(t, \xi) + \theta v_2(t, \xi)] \, d\xi.$$

Since the function $f(v, v_0) = (v - v_0 - v_0 \ln (v/v_0))/(v - v_0)^2$, $v > 0$, $v_0 > 0$, is decreasing in v for fixed v_0 and decreasing in v_0 for fixed v, and since $\alpha_k < e^{-x^*} = (\lambda - \eta)/(1 - \eta)$, there is a constant c_1 (depending only on C) such that

$$c_1 \int_{x_{k-1}}^{x_k} [w_1(t, x)]^2 \, dx \leqq V_k(t) \leqq \frac{1}{2\alpha_k} \int_{x_{k-1}}^{x_k} [w_1(t, x)]^2 \, dx. \tag{2.15}$$

We have

$$\dot{V}_k = \frac{dV_k}{dt} = \int_{x_{k-1}}^{x_k} (v_1 - e^{-x})(1 - v_1 - v_2 - \psi) \, dx$$

$$= -\int_{x_{k-1}}^{x_k} [w_1^2 + w_1(e^{-x} - 1 + \psi) + w_1 v_2] \, dx.$$

Here

$$e^{-x} - 1 + \psi = \int_0^x (v_1 - e^{-\xi} + \theta v_2) \, d\xi = \int_0^{x_{k-1}} (w_1 + \theta v_2) \, d\xi + \int_{x_{k-1}}^x (w_1 + \theta v_2) \, d\xi,$$

so that

$$w_1(e^{-x} - 1 + \psi) = (w_1 + \theta v_2) \int_{x_{k-1}}^x (w_1 + \theta v_2) \, d\xi - \theta v_2 \int_{x_{k-1}}^x (w_1 + \theta v_2) \, d\xi$$

$$+ w_1 \int_0^{x_{k-1}} (w_1 + \theta v_2) \, d\xi.$$

Since

$$\int_{x_{k-1}}^{x_k} (w_1 + \theta v_2) \left[\int_{x_{k-1}}^x (w_1 + \theta v_2) \, d\xi \right] dx = \frac{1}{2} \left[\int_{x_{k-1}}^{x_k} (w_1 + \theta v_2) \, d\xi \right]^2 \geqq 0,$$

we get for $t \geqq 0$ (using (2.5), (2.6), (2.12), and (2.15)),

$$\dot{V}_k \leqq -\int_{x_{k-1}}^{x_k} w_1^2 \, dx + \int_{x_{k-1}}^{x_k} \left[\theta v_2 \int_{x_{k-1}}^x (w_1 + \theta v_2) \, d\xi - w_1 \int_0^{x_{k-1}} (w_1 + \theta v_2) \, d\xi - w_1 v_2 \right] dx$$

$$\leqq -2\alpha_k V_k + \beta_k' \, e^{-\omega' t}[\theta(C + 1 + \theta C)\delta + C + 1]\delta + (C + 1)x^* \beta_{k-1}(e^{-\omega_{k-1} t} + \theta \, e^{-\omega' t})\delta.$$

Thus there are constants γ_k and ω'_k, where $0<\omega'_k<\min(2\alpha_k, \omega', \omega_{k-1})$, such that

$$V_k(t) \leqq \gamma_k e^{-\omega'_k t} \quad \text{for } t \geqq 0.$$

Thus, by (2.15),

$$\int_{x_{k-1}}^{x_k} |w_1(t, x)|\, dx \leqq \sqrt{\delta}\left[\int_{x_{k-1}}^{x_k} [w_1(t, x)]^2\, dx\right]^{1/2} \leqq \sqrt{\delta}\left(\frac{\gamma_k}{c_1}\right)^{1/2} e^{-(\omega'_k/2)t}.$$

For w_1 we have the equation

$$\frac{\partial w_1}{\partial t} = -v_1\left[w_1 + v_2 + \int_0^x (w_1 + \theta v_2)\, d\xi\right].$$

For $t \geqq 0$ and $x \in (x_{k-1}, x_k]$ we obtain

$$\begin{aligned} \frac{\partial w_1^2}{\partial t} &= 2w_1 \frac{\partial w_1}{\partial t} \\ &= -2v_1 w_1^2 - 2v_1 w_1\left[v_2 + \int_0^{x_{k-1}} w_1\, d\xi + \int_{x_{k-1}}^x w_1\, d\xi + \int_0^{x_{k-1}} \theta v_2\, d\xi + \int_{x_{k-1}}^x \theta v_2\, d\xi\right] \\ &\leqq -2\alpha_k w_1^2 + 2C(C+1)\Bigg[\beta'_k(1+\theta\delta)\, e^{-\omega' t} + \beta_{k-1} x^*(e^{-\omega_{k-1} t} + \theta\, e^{-\omega' t}) \\ &\qquad + \left(\frac{\delta\gamma_k}{c_1}\right)^{1/2} e^{-(\omega'_k/2)t}\Bigg]. \end{aligned}$$

There are constants β''_k and $\omega_k = \omega'_k/4$ such that

$$[w_1(t, x)]^2 \leqq (\beta''_k)^2\, e^{-2\omega_k t} \quad \text{for } t \geqq 0, x \in (x_{k-1}, x_k).$$

In the case $k=1$ this is obtained for $x=0$ also. In view of (2.12) we have thus proved (2.7), (2.8) with $\beta_k = \max(\beta_{k-1}, \beta'_k, \beta''_k)$.

After N steps we obtain

$$|v_1(t, x) - e^{-x}| \leqq \beta_N\, e^{-\omega_N t}, \qquad v_2(t, x) \leqq \beta_N\, e^{-\omega' t} \tag{2.16}$$

for $t \geqq 0$ and $x \in [0, a]$, which proves (2.1).

If T_0 is chosen so large that $\beta_N\, e^{-\omega_N T_0}$ and $\beta_N\, e^{-\omega' T_0}$ are small enough, it follows from the estimates in [1] (see in particular Remark 5.1; note a misprint at the beginning of the fourth row of Remark 5.1, where one should read $\sup_{\xi \in [0,x]} |w_i(t, \xi)|$) that

$$|v_1(t, x) - e^{-x}| \leqq \beta_0\, e^{-\nu(a)t}, \qquad v_2(t, x) \leqq \beta_0\, e^{-\nu(a)t}$$

for $t \geqq T_0, x \in [0, a]$ for a certain constant β_0, if $\nu(a) < e^{-a}$.

LEMMA 2.2. *For each $x' > x^*$ there are positive constants K and κ such that*

$$v_1(t, x) \leqq K\, e^{-\kappa t} \quad \textit{for } t \geqq 0, \quad x \in [x', L]. \tag{2.17}$$

Comment. It is first shown that $v_1 + v_2$ is bounded away from zero when x is close to x^*, and this together with Lemma 2.1 suffices to show (2.17).

Proof of Lemma 2.2. Let

$$\varphi(t) = \int_0^{x^*} [v_1(t, x) - e^{-x} + \theta v_2(t, x)]\, dx, \tag{2.18}$$

and

$$\psi_1(t, x) = \int_{x^*}^x [v_1(t, \xi) + \theta v_2(t, \xi)]\, d\xi \quad \text{for } x > x^*. \tag{2.19}$$

It follows from Lemma 2.1 (since $a<x^*$ is arbitrary) that $\varphi(t)\to 0$ as $t\to\infty$. The equations (1.1) can for $x>x^*$ be written

$$\begin{aligned}\frac{\partial v_1}{\partial t}&=v_1\left[\frac{\lambda-\eta}{1-\eta}-v_1-v_2-\varphi(t)-\psi_1(t,x)\right],\\ \frac{\partial v_2}{\partial t}&=\gamma v_2\left[\frac{\lambda-\eta}{1-\eta}-v_1-v_2-\eta\varphi(t)-\eta\psi_1(t,x)\right].\end{aligned}\tag{2.20}$$

Therefore (cf. (2.11)),

$$\frac{v_1(t,x)}{[v_2(t,x)]^{1/\gamma}}=\frac{v_1^0(x)}{[v_2^0(x)]^{1/\gamma}}\,e^{-(1-\eta)\int_0^t[\varphi(\tau)+\psi_1(\tau,x)]\,d\tau}.\tag{2.21}$$

Since $\varphi(t)\to 0$ as $t\to\infty$, we can choose T such that

$$|\varphi(t)|\leqq\frac{\lambda-\eta}{3(1-\eta)}\quad\text{for } t\geqq T.\tag{2.22}$$

Let

$$\delta_0=\frac{\lambda-\eta}{3(1-\eta)(1+\theta)C},\qquad b=x^*+\delta_0,\tag{2.23}$$

and

$$\alpha=\min\left[\alpha_0\,e^{(1-2C-(1+\theta)Cb)T},\alpha_0\,e^{\gamma(\lambda-2C-\eta(1+\theta)Cb)T},\frac{\lambda-\eta}{6(1-\eta)}\right].\tag{2.24}$$

Then it follows from (1.1) that

$$v_i(T,x)\geqq\alpha\quad\text{for } x\in(x^*,b],\quad i=1,2.$$

Suppose that $v_1(\bar t,\bar x)+v_2(\bar t,\bar x)<2\alpha$ for some $\bar t>T$ and $\bar x\in(x^*,b]$. Then there is an interval $[t',t'']$ with $t'>T$ such that $v_1(t',\bar x)+v_2(t',\bar x)=2\alpha$, and $v_1(t,\bar x)+v_2(t,\bar x)<2\alpha$ on $(t',t'']$. For $t\in(t',t'']$ and $x=\bar x$, we have, by (2.22)-(2.24),

$$\frac{\lambda-\eta}{1-\eta}-v_1-v_2-\varphi-\psi_1>\frac{\lambda-\eta}{1-\eta}-2\alpha-\frac{\lambda-\eta}{3(1-\eta)}-(1+\theta)C\delta_0\geqq 0,$$

and then (2.20) shows that $(\partial v_1/\partial t)(t,\bar x)>0$ and $(\partial v_2/\partial t)(t,\bar x)>0$ on $(t',t'']$. But this gives a contradiction, and, therefore,

$$v_1(t,x)+v_2(t,x)\geqq 2\alpha\quad\text{for all } t\geqq T,\quad x\in(x^*,b].\tag{2.25}$$

Thus

$$\psi_1(t,x)\geqq\min(1,\theta)2\alpha(x-x^*)\quad\text{for } t\geqq T,\quad x\in(x^*,b],\tag{2.26}$$

and since $\varphi(t)\to 0$ as $t\to\infty$,

$$\int_0^t[\varphi(\tau)+\psi_1(\tau,x)]\,d\tau\to\infty\quad\text{as } t\to\infty\quad\text{for } x\in(x^*,b].$$

Then we see from (2.21) that $v_1(t,x)\to 0$ as $t\to\infty$ for every $x\in(x^*,b]$. Since ψ_1 is increasing in x, (2.21) and (2.26) give an estimate of the form (2.17).

LEMMA 2.3. *For $x\in(x^*,L]$ we have that*

$$v_2(t,x)\to v_2^s(x)=e^{-x^*}\,e^{-\eta\theta(x-x^*)}\quad\textit{as } t\to\infty$$

with uniform convergence on closed subintervals of $(x^*,L]$.

Comment. First we show that $\int_{x^*}^b [v_2(t,x)-v_2^s(x)]^2\,dx \to 0$ as $t\to\infty$ for a suitable b. This is done by dividing the integral into two parts, where the first part is small because the interval is short. The other part is shown to go to zero by means of a Lyapunov type function (as in Lemma 2.1). There we use that $v_1\to 0$ and v_2 is bounded away from zero, when a neighbourhood of x^* is avoided (which follows from Lemma 2.2). After that pointwise convergence on $(x^*, b]$ is also obtained. Once we have passed the critical point x^*, we can proceed in small steps as in Lemma 2.1 and prove convergence on successively larger intervals. It may be noted that we do not obtain exponential convergence for v_2 if $x>x^*$. Indeed, experience from the linearized equation indicates that we should not expect a better rate of convergence than $1/t$.

Proof of Lemma 2.3. Let b and α be as in (2.23) and (2.24), and let $\varepsilon>0$ be given. Choose $x'\in(x^*, b)$ such that

$$(2.27)\qquad (C+e^{-x^*})^2(x'-x^*)<\frac{\varepsilon}{2}\quad\text{and}\quad \eta\theta(C+e^{-x^*})^2(b-x')(x'-x^*)<\frac{\alpha\varepsilon c_1}{4},$$

where c_1 is the constant in (2.15). Consider the function

$$V(t)=\int_{x'}^b\left[v_2(t,x)-v_2^s(x)-v_2^s(x)\ln\frac{v_2(t,x)}{v_2^s(x)}\right]dx,$$

where $v_2^s(x)$ is given by (1.8). Since $v_1(t,x)\to 0$ as $t\to\infty$ uniformly on $[x', b]$ (Lemma 2.2), it follows from (2.25) that there is a t_1 (depending on x', hence on ε) such that

$$v_2(t,x)\geqq\alpha\quad\text{for } t\geqq t_1,\quad x\in[x', b].$$

From the properties of the function $(v-v_0-v_0\ln(v/v_0))/(v-v_0)^2$ we find as in (2.15) (since $v_2^s(b)>\alpha$) that

$$(2.28)\qquad c_1\int_{x'}^b[w_2(t,x)]^2\,dx\leqq V(t)\leqq\frac{1}{2\alpha}\int_{x'}^b[w_2(t,x)]^2\,dx\quad\text{for } t\geqq t_1,$$

where

$$w_2(t,x)=v_2(t,x)-v_2^s(x).$$

The derivative of V is

$$\begin{aligned}\dot V&=\int_{x'}^b\gamma(v_2-v_2^s)\left(e^{-x^*}-v_1-v_2-\eta\varphi-\eta\int_{x^*}^x v_1\,d\xi-\eta\theta\int_{x^*}^x v_2\,d\xi\right)dx\\&=\gamma\int_{x'}^b w_2\left(-w_2-\eta\theta\int_{x^*}^x w_2\,d\xi-\varphi_1\right)dx,\end{aligned}$$

where φ is defined in (2.18), and

$$\varphi_1(t,x)=v_1(t,x)+\eta\varphi(t)+\eta\int_{x^*}^x v_1(t,\xi)\,d\xi,$$

so that $\varphi_1(t,x)\to 0$ as $t\to\infty$, uniformly on $[x', b]$. Since

$$\int_{x'}^b w_2(t,x)\left[\int_{x'}^x w_2(t,\xi)\,d\xi\right]dx=\frac{1}{2}\left[\int_{x'}^b w_2(t,\xi)\,d\xi\right]^2\geqq 0,$$

we have

$$\dot V(t)\leqq-\gamma\int_{x'}^b[w_2(t,x)]^2\,dx-\gamma\int_{x'}^b w_2(t,x)\left[\eta\theta\int_{x^*}^{x'}w_2(t,\xi)\,d\xi+\varphi_1(t,x)\right]dx.$$

There is a $t_2 \geqq t_1$ such that

$$|\varphi_1(t, x)| < \frac{\alpha\varepsilon c_1}{4(C+e^{-x^*})(b-x')} \quad \text{for } t \geqq t_2 \quad \text{and} \quad x \in [x', b].$$

For $t \geqq t_2$ we get, from (2.28) and (2.27),

$$\dot{V}(t) \leqq -2\alpha\gamma V(t) + \gamma(C+e^{-x^*})\left[\eta\theta(C+e^{-x^*})(x'-x^*) + \frac{\alpha\varepsilon c_1}{4(C+e^{-x^*})(b-x')}\right](b-x')$$

$$\leqq -2\alpha\gamma V(t) + \frac{1}{2}\gamma\alpha\varepsilon c_1.$$

Thus

$$V(t) \leqq V(t_2)\, e^{-2\alpha\gamma(t-t_2)} + \tfrac{1}{4}\varepsilon c_1 \quad \text{for } t \geqq t_2,$$

and there is a $t_3 \geqq t_2$ such that

$$V(t) \leqq \tfrac{1}{2}\varepsilon c_1 \quad \text{for } t \geqq t_3.$$

Consequently, by (2.27) and (2.28),

$$\int_{x^*}^{b} [w_2(t, x)]^2\, dx = \int_{x^*}^{x'} w_2^2\, dx + \int_{x'}^{b} w_2^2\, dx < (C+e^{-x^*})^2(x'-x^*) + \frac{1}{c_1} V(t)$$

$$\leqq \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon \quad \text{for } t \geqq t_3 = t_3(\varepsilon).$$

This proves that

$$\int_{x^*}^{b} [w_2(t, x)]^2\, dx \to 0 \quad \text{as } t \to \infty,$$

and by Cauchy-Schwarz's inequality,

$$\text{(2.29)} \qquad \int_{x^*}^{b} |w_2(t, x)|\, dx \to 0 \quad \text{as } t \to \infty.$$

Fix $x \in (x^*, b]$. Again it follows from (2.25) that there is a t_4 such that $v_2(t, x) \geqq \alpha$ for $t \geqq t_4$. For $t \geqq t_4$ we have

$$\frac{\partial w_2^2}{\partial t} = 2w_2 \frac{\partial w_2}{\partial t} = -2\gamma w_2 v_2 \left(w_2 + \eta\theta \int_{x^*}^{x} w_2\, d\xi + \varphi_1 \right)$$

$$\leqq -2\alpha\gamma w_2^2 + \varphi_2,$$

where

$$\varphi_2(t, x) = -2\gamma v_2(t, x) w_2(t, x)\left[\eta\theta \int_{x^*}^{x} w_2(t, \xi)\, d\xi + \varphi_1(t, x)\right],$$

and it follows from (2.29) that $\varphi_2(t, x) \to 0$ as $t \to \infty$. Thus

$$[w_2(t, x)]^2 \leqq [w_2(t_4, x)]^2\, e^{-2\alpha\gamma(t-t_4)} + \int_{t_4}^{t} e^{-2\alpha\gamma(t-\tau)}\, \varphi_2(\tau, x)\, d\tau \quad \text{for } t \geqq t_4,$$

so that $w_2(t, x) \to 0$ as $t \to \infty$. The convergence is uniform on closed subintervals of $(x^*, b]$.

For the last part of the proof we divide $[b, L]$ into N intervals $[x_{k-1}, x_k]$, $k = 1, \dots, N$, of length

$$x_k - x_{k-1} \leqq \frac{v_2^s(L)}{4\eta\theta C}.$$

Assume that it is proved that $w_2(t, x) \to 0$ as $t \to \infty$ for $x \in (x^*, x_{k-1}]$ (this is the case for $k = 1$). Choose T_k such that

$$|\varphi_1(t, x)| \leqq \tfrac{1}{4}\, v_2^s(L) \quad \text{for } x \in (x_{k-1}, x_k], \quad t \geqq T_k,$$

and

$$\eta\theta \int_{x^*}^{x_{k-1}} |w_2(t, \xi)|\, d\xi \leqq \frac{1}{4}\, v_2^s(L) \quad \text{for } t \geqq T_k.$$

Let

$$\alpha_k = \min\,[\tfrac{1}{4}\, v_2^s(L), \alpha_0\, e^{\gamma(\lambda - 2C - \eta(1+\theta)CL)T_k}].$$

(Note that x_k, T_k, and α_k are not the same as in Lemma 2.1, although the usage is similar.) For $t \geqq T_k$ and $x \in (x_{k-1}, x_k]$, we have

$$\begin{aligned} &e^{-x^*} - v_1(t, x) - v_2(t, x) - \eta\varphi(t) - \eta\psi_1(t, x) \\ &= v_2^s(x_{k-1}) - \varphi_1(t, x) - v_2(t, x) - \eta\theta \int_{x^*}^{x_{k-1}} w_2(t, \xi)\, d\xi - \eta\theta \int_{x_{k-1}}^{x} v_2(t, \xi)\, d\xi \\ &\geqq v_2^s(L) - \frac{3}{4}\, v_2^s(L) - v_2(t, x) = \frac{1}{4}\, v_2^s(L) - v_2(t, x), \end{aligned}$$

so that $(\partial v_2/\partial t)(t, x) \geqq 0$, if $v_2(t, x) \leqq \alpha_k$. Since $v_2(T_k, x) \geqq \alpha_k$, this implies that $v_2(t, x) \geqq \alpha_k$ for all $t \geqq T_k$, $x \in (x_{k-1}, x_k]$. By means of the function

$$V_k(t) = \int_{x_{k-1}}^{x_k} \left[v_2(t, x) - v_2^s(x) - v_2^s(x) \ln \frac{v_2(t, x)}{v_2^s(x)} \right] dx,$$

we obtain just as before (when we used $V(t)$ to show that $w_2(t, x) \to 0$ as $t \to \infty$ for $x \in (x^*, b]$) that $\int_{x_{k-1}}^{x_k} |w_2(t, x)| dx \to 0$ as $t \to \infty$, and then $w_2(t, x) \to 0$ as $t \to \infty$ for $x \in (x_{k-1}, x_k]$. In this way we obtain after N steps that $w_2(t, x) \to 0$ as $t \to \infty$ for all $x \in (x^*, L]$. The convergence is uniform on closed subintervals of $(x^*, L]$.

By this, Theorem 1 is proved.

3. Some concluding remarks. Let us remark what happens for parameter values that do not satisfy (1.5), (1.6). If $\eta < \lambda < 1$, but $x^* \geqq L$, then there is only one zone, and convergence follows from Lemma 2.1. If $\lambda < 1$ and $\eta \geqq \lambda$, then there is no critical point x^*, and again there is only one zone. We can apply Lemma 2.1 with $a = L$; we replace x^* by L and $(\lambda - \eta)/(1 - \eta)$ by e^{-L}. If $\lambda = 1$ and $\eta < 1$, then we can prove Lemmas 2.2 and 2.3 with $x^* = 0$. If $\eta > \lambda > 1$, or $\lambda > 1$ and $0 < \eta \leqq \lambda$, or $\lambda = 1$ and $\eta > 1$, then the result follows by symmetry. Indeed, if we make the substitutions $\bar{t} = \lambda\gamma t$, $\bar{x} = \eta\theta x$, $\bar{v}_i = v_i/\lambda$, $\bar{\lambda} = 1/\lambda$, $\bar{\gamma} = 1/\gamma$, $\bar{\eta} = 1/\eta$, and $\bar{\theta} = 1/\theta$, then we obtain the equations (1.1) with all quantities barred and with indices 1 and 2 interchanged, and we can apply the previous results. The case $\eta = 0$ is also easily treated.

The case $\lambda=\eta=1$ is exceptional, as is indicated in [1], and requires a special treatment. Following [1] we define $v_1=F(x,u)\geqq 0$ as the solution of

$$v_1+\zeta(x)v_1^\gamma=u, \qquad u\geqq 0,$$

where $\zeta(x)=v_2^0(x)/[v_1^0(x)]^\gamma$. Define $G(x,u)=u-F(x,u)$, $F_\gamma(x,u)=F(x,u)+\gamma G(x,u)$, and $F_\theta(x,u)$ analogously. Then, as is shown in [1], the system (1.1) with $\lambda=\eta=1$ is equivalent to the single equation

$$\frac{\partial u}{\partial t}=F_\gamma(x,u)\left[1-u-\int_0^x F_\theta(\xi,u)\,d\xi\right], \tag{3.1}$$

with $v_1(t,x)=F(x,u(t,x))$ and $v_2(t,x)=G(x,u(t,x))$, or $u(t,x)=v_1(t,x)+v_2(t,x)$. It is shown in [1] that

$$1-u(x)-\int_0^x F_\theta(\xi,u(\xi))\,d\xi=0 \tag{3.2}$$

has a unique solution $u^s(x)$, and we can now prove that $u(t,x)\to u^s(x)$ as $t\to\infty$ with exponential convergence. We omit the details and only indicate the main steps of the proof.

Divide $[0,L]$ into N intervals $[x_{k-1},x_k]$ of length δ, where δ is sufficiently small. Let $w(t,x)=u(t,x)-u^s(x)$, and assume that

$$|w(t,x)|\leqq \beta_{k-1}\,e^{-\omega_{k-1}t}, \qquad t\geqq 0, x\in[0,x_{k-1}] \tag{3.3}$$

for some positive constants β_{k-1} and ω_{k-1}. Much as in Lemma 2.1 we can show that

$$u(t,x)\geqq \alpha_k>0, \quad t\geqq 0, \quad x\in(x_{k-1},x_k]. \tag{3.4}$$

Introduce the function

$$V_k(t)=\int_{x_{k-1}}^{x_k}\int_{u^s(x)}^{u(t,x)}\frac{F_\theta(x,\bar u)-F_\theta(x,u^s(x))}{F_\gamma(x,\bar u)}\,d\bar u\,dx. \tag{3.5}$$

From the properties of $F(x,u)$ and (3.4) we obtain

$$a_k\int_{x_{k-1}}^{x_k}[w(t,x)]^2\,dx\leqq V_k(t)\leqq b_k\int_{x_{k-1}}^{x_k}[w(t,x)]^2\,dx \tag{3.6}$$

for certain positive constants a_k and b_k. From (3.1) and (3.2) we get

$$\frac{\partial w}{\partial t}=F_\gamma(x,u)\left[-w-\int_0^x \Delta(t,\xi)\,d\xi\right], \tag{3.7}$$

where

$$\Delta(t,x)=F_\theta(x,u(t,x))-F_\theta(x,u^s(x)), \tag{3.8}$$

so that

$$\Delta(t,x)w(t,x)\geqq c_k[w(t,x)]^2, \quad t\geqq 0, \quad x\in(x_{k-1},x_k] \tag{3.9}$$

for some positive constant c_k. Then

$$\dot V_k(t)=\int_{x_{k-1}}^{x_k}\frac{\Delta(t,x)}{F_\gamma(x,u)}\frac{\partial w}{\partial t}\,dx=-\int_{x_{k-1}}^{x_k}\Delta(t,x)w(t,x)\,dx$$
$$-\int_{x_{k-1}}^{x_k}\Delta(t,x)\int_0^{x_{k-1}}\Delta(t,\xi)\,d\xi\,dx-\int_{x_{k-1}}^{x_k}\Delta(t,x)\int_{x_{k-1}}^{x}\Delta(t,\xi)\,d\xi\,dx.$$

The important thing is that the last term can be estimated:

$$\int_{x_{k-1}}^{x_k} \Delta(t, x) \int_{x_{k-1}}^{x} \Delta(t, \xi)\, d\xi\, dx = \frac{1}{2}\left(\int_{x_{k-1}}^{x_k} \Delta(t, x)\, dx\right)^2 \geqq 0.$$

The first term can be estimated in terms of $V_k(t)$ by (3.6) and (3.9), and the middle term tends exponentially to zero as $t \to \infty$, according to (3.3) (since $|\Delta| \leqq (1+\theta)|w|$). Thus $V_k(t)$, and hence $\int_{x_{k-1}}^{x_k} [w(t, x)]^2\, dx$, tends exponentially to zero. Finally, a pointwise estimate for $[w(t, x)]^2$, if $t \geqq 0$, $x \in (x_{k-1}, x_k]$, is obtained from (3.7) as in the proof of Lemma 2.1.

In this way we know the asymptotic behaviour of the solution of (1.1) for all values of the parameters.

Acknowledgment. The author wishes to thank L. Bass, A. M. Fink, and A. McNabb for many helpful discussions on this problem.

REFERENCES

[1] L. Bass, A. J. Bracken, K. Holmåker, and B. R. F. Jefferies, *Integro-differential equations for the self-organisation of liver zones by competitive exclusion of cell-types*, J. Austral. Math. Soc. Ser. B, 29 (1987), pp. 156-194.

[2] L. Bass, A. J. Bracken, and R. Vyborny, *Minimisation problems for implicit functionals defined by differential equations of liver kinetics*, J. Austral. Math. Soc. Ser. B, 25 (1984), pp. 538-562.

[3] A. M. Fink, *Optimal control in liver kinetics*, J. Austral. Math. Soc. Ser. B, 27 (1986), pp. 361-369.

[4] K. Holmåker, *An optimal control problem in the study of liver kinetics*, J. Optim. Theory Appl., 48 (1986), pp. 289-302.

[5] K. Holmåker and D. Stewart, *A class of optimization problems with noncompact constraints: general results and applications*, SIAM J. Control Optim., 25 (1987), pp. 1032-1052.

SIAM J. MATH. ANAL.
Vol. 24, No. 1, pp. 129–154, January 1993

REMARKS ON PERIODS OF PLANAR HAMILTONIAN SYSTEMS*

FRANZ ROTHE[†]

In memoriam of Professor Peter Henrici for all his help.

Abstract. Criteria for monotonicity of the periods of planar Hamiltonian systems from various authors are logically related. From the more restrictive criteria, not only monotonicity is proved, but even a further convexity property of the energy-period function. This implies estimates for the canonical partition function—which is the Laplace transform of the energy-period function—and some averages from the canonical ensemble of thermodynamics. For the standard system of classical mechanics with quadratic kinetic energy and symmetric potential energy, the Laplace transform technique allows one to solve the inverse problem to determine the potential energy function from given periods. Several examples are discussed.

Key words. energy-period function, monotonicity, Laplace transform

AMS(MOS) subject classifications. 34C25, 34A25, 34A55

1. Introduction. In [17], Waldvogel proves that the oscillation periods of the well-known Volterra–Lotka system

$$\begin{aligned} \frac{du}{dt} &= \lambda u\,(1-v)\,, \\ \frac{dv}{dt} &= -\mu v\,(1-u) \end{aligned} \tag{1}$$

increase monotonically with the amplitude of the oscillations. A different variant of the proof based on the canonical ensemble from classical equilibrium thermodynamics is given by Rothe [12].

Introducing the variables (p,q) via $u = e^p$ and $v = e^q$ and exchanging the two equations transforms (1) into the Hamiltonian system,

$$\frac{dq}{dt} = \frac{\partial H}{\partial p}\,, \qquad \frac{dp}{dt} = -\frac{\partial H}{\partial q} \tag{2}$$

with the Hamiltonian

$$H(p,q) = \lambda\,(e^q - q - 1) + \mu\,(e^p - p - 1)\,. \tag{3}$$

In this note, we generalize the Hamiltonian (3) from the Volterra–Lotka system and treat the case of arbitrary Hamiltonians with separated variables

$$H(p,q) = \lambda F(q) + \mu G(p). \tag{4}$$

For the functions $F = F(q)$ and $G = G(p)$, we assume throughout that

- $(F0)$ $F \in C^4(\mathbb{R},\mathbb{R})$ is at least four times continuously differentiable, $F(0) = 0$; and
- $(f0)$ The derivative $f = F'$ satisfies $f(0) = 0$, $f'(0) > 0$, and $qf(q) > 0$ for all $q \neq 0$

*Received by the editors March 8, 1991; accepted for publication (in revised form) March 11, 1992.

[†]Mathematics Department, University of North Carolina, Charlotte, North Carolina 28223.

and similarly for the function $G = G(q)$ and its derivative $g = G'$. Thus we treat a system

$$\frac{dq}{dt} = \mu g(p), \qquad \frac{dp}{dt} = -\lambda f(q), \tag{5}$$

with $\lambda > 0$ and $\mu > 0$ and $(f0)$ holding for both functions f and g. Especially, our class of systems includes the standard example from classical mechanics

$$\frac{dq}{dt} = p, \qquad \frac{dp}{dt} = -f(q), \tag{6}$$

which results from (5) by setting $\lambda = \mu = 1$, $G(p) = \frac{1}{2}p^2$ for the kinetic energy and keeping only the one potential $F = F(q)$.

Systems (5) and (6) have a unique center at $(p, q) = (0, 0)$, which is surrounded by periodic orbits $(t, E, \lambda, \mu) \in \mathrm{IR}\times[0, \infty)\times(0, \infty)^2 \mapsto (p, q) = (p, q)(t, E, \lambda, \mu) \in \mathrm{IR}^2$. We are interested in estimates for the *energy-period function* $(E, \lambda, \mu) \in [0, \infty)\times(0, \infty)^2 \mapsto T = T(t, E, \lambda, \mu) \in (0, \infty)$, where T is the primitive period of the oscillations of system (5) given by

$$\begin{aligned} p(t + T(E, \lambda, \mu), E, \lambda, \mu) &= p(t, E, \lambda, \mu), \\ q(t + T(E, \lambda, \mu), E, \lambda, \mu) &= q(t, E, \lambda, \mu), \\ H(p(t, E, \lambda, \mu), q(t, E, \lambda, \mu)) &= E \\ \text{for all } t \in \mathrm{IR},\ E \in [0, \infty),\ &\lambda, \mu \in (0, \infty). \end{aligned} \tag{7}$$

In this introduction, we quote a typical new result of this note. Then we explain the application to Hamiltonian systems of several degress of freedom. Next, we indicate the connection with equilibrium thermodynamics. Finally, we comment on the application to bifurcation of two point boundary value problems.

Assume that both functions f and g satisfy

$$\begin{aligned} f'(q) > 0 &\text{ for all } q \in \mathrm{IR}, \\ q\left[f''(q) - \frac{f''(0)}{f'^2(0)}\left(f'^2(q) - \frac{1}{3}f(q)f''(q)\right)\right] < 0 &\text{ for all } q \neq 0. \end{aligned} \tag{8}$$

Then the energy-period function $T = T(E, \lambda, \mu)$ is a strictly increasing function of the energy E. The apparently much simpler assumption $qf''(q) < 0$ for all $q \neq 0$ is sufficient for (8) to hold. But since it implies $f''(0) = 0$, it is often too restrictive.

The following remark is based on work of Arnold [1]. In the theory of Hamiltonian dynamics, it is a basic problem to study the effects introduced by a weak coupling into systems of several independent oscillators. As the simplest, really hard example, take a system of two degrees of freedom with Hamiltonian

$$H = H_0 + \varepsilon H_1, \tag{9}$$

where H_0 is the Hamiltonian of the uncoupled system, H_1 is the coupling, and ε is a small coupling constant. Suppose that the unperturbed system is completely integrable and its Hamiltonian is

$$H_0 = H_{01}(J_1) + H_{02}(J_2), \tag{10}$$

where J_1 and J_2 are the action variables for the first and second degree of freedom.

Arnold's celebrated result from [1] states that under the appropriate *nondegeneracy assumption,* in the weak coupling limit $\varepsilon \to 0$ a part of the phase space Γ of nearly full Lesbesgue-measure is filled with nonresonant KAM-TORI. Indeed, Arnold gives results in two different settings, for which the relevant phase space Γ is

(a) $\Gamma = \{ (p_1, q_1, p_2, q_2) \} = \mathbb{R}^4$, the entire four-dimensional space;

(b) $\Gamma = \{ (p_1, q_1, p_2, q_2) \mid H(p_1, q_1, p_2, q_2) = E \}$, a three-dimensional surface with specified total energy E.

The nondegeneracy assumptions appropriate for case (a) and (b) are

$$\text{(11a)} \qquad \frac{\partial^2 H_{01}}{\partial J_1^2} \frac{\partial^2 H_{02}}{\partial J_2^2} \neq 0,$$

$$\text{(11b)} \qquad \frac{\partial^2 H_{01}}{\partial J_1^2} \left(\frac{\partial H_{02}}{\partial J_2} \right)^2 + \frac{\partial^2 H_{02}}{\partial J_2^2} \left(\frac{\partial H_{01}}{\partial J_1} \right)^2 \neq 0.$$

Assumptions (11a) and (11b) can be expressed in terms of the energy-period function instead of the action-angle variables. We consider the Hamiltonian system (2) with one degree of freedom dropping the index 1 or 2. The solutions of that system as specified by (7) give rise to a mapping $(t, E) \mapsto (p, q)$. We substitute $(p, q) = (p, q)(t, E)$ into the integral for the area $A(E)$ enclosed by the orbit $(p, q)(t, E)$. The determinant of the Jacobian of this mapping is equal to one, and hence $dp\,dq = dtdE$. Thus we get

$$\text{(12)} \qquad A(E) = \int_{\{(p,q) | H(p,q) \leq E\}} dp\,dq = \int_0^E T(\widetilde{E})\, d\widetilde{E},$$

and hence $dA/dE = T$. In action-angle variables (J, ϕ), the Hamiltonian $H = H(J)$ depends only on the action J, and the equations of motion are

$$\text{(13)} \qquad \frac{dJ}{dt} = 0, \qquad \frac{d\phi}{dt} = \frac{dH}{dJ}.$$

From the obvious solution $\phi(t) = (dH/dJ)\, t + \text{const}$, we see that $dH/dJ = 2\pi/T(H)$ for the orbit with $H = E$. Hence

$$\text{(14)} \qquad \frac{dH}{dJ} T = \frac{dH}{dJ} \frac{dA}{dH} = 2\pi,$$

which shows that the Hamiltonian $H = H(J)$ has the inverse function $J = A(H)/(2\pi)$. Calculating the second derivative for an inverse function, we get

$$\text{(15)} \qquad \frac{d^2 H}{dJ^2} = -\frac{d^2 J}{dH^2} \left(\frac{dJ}{dH} \right)^{-3} = -\frac{4\pi^2}{T^3} \frac{dT}{dH}.$$

Hence the nondegeneracy assumptions can be expressed in terms of the energy-period function $T = T(E)$ as

$$\text{(16a)} \qquad \frac{dT(E_1)}{dE} \frac{dT(E - E_1)}{dE} \neq 0,$$

$$\text{(16b)} \qquad \frac{d \log T(E_1)}{dE} + \frac{d \log T(E - E_1)}{dE} \neq 0.$$

For the nondegeneracy assumption (16a) to hold for all E and E_1, it is obviously necessary that the energy-period function is strictly increasing or decreasing. For many examples, this can be checked by means of the criteria from this note. The isoenergetical nondegeneracy assumption states that for fixed total energy $E = E_1 + E_2$, the frequency ratio $T(E_1)/T(E - E_1)$ of the two uncoupled subsystems with index 1 and 2 has a nonvanishing derivative with respect to E_1.

Our second remark takes the point of view of classical equilibrium thermodynamics. Assume that system (2) is coupled to a large heat bath with inverse absolute temperature β. All quantities of equilibrium thermodynamics can be derived from the *canonical partition function* $Z(\beta)$ defined as

$$Z(\beta) = \int \exp\left(-\beta H(p,q)\right)\, dp\, dq. \tag{17}$$

For example, the mean value and variance of the energy $H(p,q)$ are

$$\langle H\rangle_{\text{can}} = -\frac{\partial \log Z(\beta)}{\partial \beta} \quad \text{and} \quad \text{Var}_{\text{can}} H = \frac{\partial^2 \log Z(\beta)}{\partial \beta^2}, \tag{18}$$

and the specific heat is $\beta^2 \text{Var}_{\text{can}} H$. (See, e.g., Becker [2, p. 127].)

We substitute the solutions $(p,q) = (p,q)(t,E)$ of system (2) into definition (17). Because the mapping $(t,E) \mapsto (p,q)$ is area preserving, we get

$$Z(\beta,\lambda,\mu) = \int_0^\infty e^{-\beta E} T(E,\lambda,\mu)\, dE, \tag{19}$$

which shows that the canonical partition function is the Laplace transform of the energy-period function. This connection allows us to get bounds for the mean energy and the specific heat as well. We quote a typical result: If assumption (8) holds for both functions f and g, then the mean energy is bounded below by $\langle H\rangle_{\text{can}} > \beta^{-1}$. This lower bound is the actual value for the harmonic oscillator.

Several authors (see Schaaf [15] for a survey and an extensive biography) have considered the energy-period function in order to find or exclude bending points in the bifurcation diagram of two point boundary value problems. The most simple example is the nonlinear elliptic Dirichlet boundary value problem

$$\begin{aligned} &u_{xx} + f(u) = 0 \quad \text{for all } x \in [0,T], \\ &u(0) = u(T) = 0. \end{aligned} \tag{20}$$

In the bifurcation diagram, the bifurcation parameter T is plotted against a norm of u. The differential equation in (20) is equivalent to a Hamiltonian system with the variables $q = u$ and $p = u_x$. The energy $E = u_x^2/2 + F(u)$ serves as a norm of u.

The *second* branch in the bifurcation diagram—corresponding to solutions of (20) with one sign change for $x \in (0,T)$—is just the plot of the energy-period function $T = T(E)$ with the axes interchanged. The part of the *first* branch corresponding to positive solutions of (20) comes from the plot of the energy-halfperiod function $T = T^+(E)$, defined via

(21)

$$\begin{aligned} q(T^+(E),E) = q(0,E) &= 0, \\ q(u,E) \geq 0 \quad \text{and} \quad \tfrac{1}{2}p^2(u,E) + F(q(u,E))) &= E \\ &\text{for all } E \in [0,\infty) \quad \text{and} \quad u \in [0,T^+(E)]. \end{aligned}$$

The other part of the first branch corresponds to negative solutions, for which we define $T^- = T^-(E)$ by reversing the sign of q in (21).

Satisfactory monotonicity criteria for the first branch are given by Chafee [5], Opial [11], Schaaf [15], and others. We mention them briefly in Proposition 4 below. The situation is much more complicated for the second branch. It is not always possible to conclude monotonicity of the second branch by applying the same criteria as for the first branch. Indeed, for many nonsymmetric systems with $f(-u) \neq -f(u)$, as, e.g., (20) with $f(u) = e^u - 1$, monotonicity of the period $T = T^+(E) + T^-(E)$ results after the cancelling of an increasing halfperiod T^- and a decreasing halfperiod T^+.

In this note, we give several criteria for monotonicity of the entire period $T = T(E)$. This corresponds to the second bifurcation branch of (20), which brings in additional difficulties. Different critera deal with them in different manners. We can aim at estimates for the second derivative $T''(E)$ as Schaaf in [14] and a variety of results in [15], and Chicone [3] have done. Or we can try to eliminate or reduce the asymmetry by substraction of suitable terms (e.g., linear terms of the force). This is the approach of Chow and Wang [4] and several new criteria in this note.

The present note assembles these results and states implications among them. Thus we prove that our new monotonicity criteria $f \in \mathfrak{f}_4^\pm$ or $F \in \mathfrak{F}_4^\pm$ are more general than Schaaf's criterion $f \in \mathfrak{f}_1^\pm$. But in concrete examples, the latter turns out to be the most useful criterion. Fortunately, we get some new results from the more stringent criteria involving the second derivative, too: new convexity properties of the energy-period functions and estimates of thermodynamic averages.

2. Notation and main results. We define some sets of functions appearing as right-hand sides of system (5) and as primitive functions of them, respectively.

DEFINITION 1. By $\mathfrak{f} \subset C^3(\mathbb{R}, \mathbb{R})$ we denote the set of all functions satisfying

$$(f0) \qquad f(0) = 0,\ f'(0) > 0 \ \text{ and } \ qf(q) > 0 \quad \text{for all } \ q \neq 0.$$

By $\mathfrak{n} \subset \mathfrak{f}$ we denote the set of all functions for which there exist $q_1 \in [-\infty, 0)$, $q_2 \in (0, \infty]$ such that

$$(n) \qquad \begin{aligned} f'(q) > 0 &\quad \text{for all } \ q \in (q_1, q_2)\,, \\ f'(q) < 0 &\quad \text{for all } \ q \in \mathbb{R} \setminus [q_1, q_2]\,. \end{aligned}$$

By $\mathfrak{m} \subset \mathfrak{n}$ we denote the set of all functions satisfying

$$(m) \qquad f(0) = 0 \ \text{ and } \ f'(q) > 0 \quad \text{for all } \ q \neq 0.$$

By $\mathfrak{F} \subset C^4(\mathbb{R}, \mathbb{R})$ we denote the set of all functions satisfying

$$(F0) \qquad F(0) = 0 \ \text{ and } \ F' \in \mathfrak{f}\,.$$

By $\mathfrak{M} \subset \mathfrak{F}$ we denote the set of all functions satisfying

$$(M) \qquad F(0) = 0 \ \text{ and } \ F' \in \mathfrak{m}\,.$$

We denote the differentiation operator by D. Obviously, $\mathfrak{m} \subset \mathfrak{n} \subset \mathfrak{f} = D\mathfrak{F}$ and $\mathfrak{m} = D\mathfrak{M}$. Next we define some sets of functions f and g giving rise to a monotonic energy-period function. We need the following functions:

$$(22) \qquad \begin{aligned} f_1 &\equiv 5f''^2 - 3f'f''', \\ f_4 &\equiv q\,[f_0''\,(3f'^2 - ff'') \, - 3f_0'^2 f''] \,, \end{aligned}$$

corresponding to the function f. The index zero denotes values at $q = 0$.

Definition 2. We define classes of functions related to increasing periods. By $\mathfrak{f}_1^+ \subset \mathfrak{m}$ we denote the set of all functions satisfying

$$(f1)^+ \qquad f'(q) > 0 \text{ and } f_1(q) > 0 \quad \text{for all } q \in \mathbb{R}.$$

By $\mathfrak{f}_2^+ \subset \mathfrak{n}$ we denote the set of all functions satisfying

$$(f2)^+ \qquad \begin{aligned} &f'(q) > 0 \text{ and } f_1(q) > 0 && \text{for all } q \in (q_1, q_2)\,, \\ &f'(q) < 0 \text{ and } f_4(q) > 0 && \text{for all } q \in \mathbb{R} \setminus [q_1, q_2]\,. \end{aligned}$$

By $\mathfrak{f}_3^+ \subset \mathfrak{n}$ we denote the set of all functions satisfying

$$(f3)^+ \qquad \begin{aligned} &f'(q) > 0 \text{ and } f_1(q) > 0 && \text{for all } q \in (q_1, q_2)\,, \\ &f'(q) < 0 && \text{for all } q \in \mathbb{R} \setminus [q_1, q_2]\,. \end{aligned}$$

By $\mathfrak{f}_4^+ \subset \mathfrak{n}$ we denote the set of all functions satisfying

$$(f4)^+ \qquad \begin{aligned} &f'(q) > 0 \text{ and } f_4(q) > 0 && \text{for all } q \in (q_1, 0) \cup (0, q_2)\,, \\ &f'(q) < 0 && \text{for all } q \in \mathbb{R} \setminus [q_1, q_2]\,. \end{aligned}$$

The classes $(\overline{f1})^+$ through $(\overline{f4})^+$ of functions corresponding to nondecreasing periods are defined by replacing in $(f1)^+$ through $(f4)^+$ the assumptions $f_1(q) > 0$ and $f_4(q) > 0$ by $f_1(q) \geq 0$ and $f_4(q) \geq 0$.

Definition 3. We define the following classes of functions related to decreasing periods. For $i = 1$ and $i = 4$ we denote by $\mathfrak{f}_i^- \subset \mathfrak{m}$ the set of all functions satisfying

$$(fi)^- \qquad f'(q) > 0 \quad \text{for all } q \in \mathbb{R} \text{ and } f_i(q) < 0 \quad \text{for all } q \neq 0.$$

For $i = 1$ and $i = 4$ we denote by $\bar{\mathfrak{f}}_i^- \subset \mathfrak{m}$ the set of all functions satisfying

$$(\overline{fi})^- \qquad f'(q) > 0 \quad \text{for all } q \in \mathbb{R} \text{ and } f_i(q) \leq 0 \quad \text{for all } q \in \mathbb{R}.$$

Corresponding to the potential F, we define the functions

$$\begin{aligned} F_1 &\equiv 2F\,(3f'^2 - ff'') \,-\, 3f^2 f'\,, \\ F_2 &\equiv \left[1 + \tfrac{1}{3} f_0'' f_0'^{-2} f\right] \left[F\,(3f'^2 - ff'') \,-\, f^2 f'\right] \,-\, F f'^2\,, \\ F_3 &\equiv f_0'^{-1/2} \,-\, (2F)^{3/2} f^{-3} f'\,, \\ F_4 &\equiv 3 f_0'^2\,(f^2 - 2Ff') \,+\, f_0'' f^3\,, \\ F_5 &\equiv f_0'^2 f^6 f' F_4 + (2F)^{3/2} F_1\,. \end{aligned} \tag{23}$$

Definition 4. The following classes of potentials $F = F(q)$ correspond to monotonic periods. For $i = 1, 2, 3, 4$, we denote by $\mathfrak{F}_i^+ \subset \mathfrak{F}$ the set of all functions satisfying

$$(Fi)^+ \qquad F_i(q) > 0 \quad \text{for all } q \neq 0\,.$$

By $\mathfrak{F}_5^+ \subset \mathfrak{F}$ we denote the set of all functions for which there exist $q_1 \in [-\infty, 0)$, $q_2 \in (0, \infty]$ such that $f = F'$ satisfies

$$(F5)^+ \qquad \begin{aligned} &f'(q) > 0 \text{ and } F_5(q) > 0 && \text{for all } q \in (q_1, 0) \cup (0, q_2)\,, \\ &f'(q) < 0 && \text{for all } q \in \mathbb{R} \setminus [q_1, q_2]\,. \end{aligned}$$

For $i = 1, 2, 3, 4$, we denote by $\mathfrak{F}_i^- \subset \mathfrak{F}$ the set of all functions satisfying

$$(Fi)^- \qquad F_i(q) < 0 \quad \text{for all } q \neq 0\,.$$

The classes $(\overline{Fi})^-$ for $i = 1, 2, 3, 4$ are defined by including the equality sign in the assumptions for $F_i(q)$.

DEFINITION 5. The following classes of functions are related to increasing half-periods. By $\mathfrak{h}_1^+ \subset \mathfrak{f}$ we denote the set of all functions in $\mathfrak{f}$ such that

$$(h1)^+ \qquad qf''(q) < 0 \quad \text{for all } q \neq 0 .$$

By $\mathfrak{H}_1^+ \subset \mathfrak{F}$ we denote the set of all functions in $\mathfrak{F}$ such that

$$(H1)^+ \qquad f^2 - 2Ff' > 0 \quad \text{for all } q \neq 0 .$$

By $\mathfrak{h}_2^+ \subset \mathfrak{f}$ we denote the set of all functions in $\mathfrak{f}$ such that

$$(h2)^+ \qquad q\frac{d}{dq}\frac{f(q)}{q} < 0 \quad \text{for all } q \neq 0 .$$

Finally, let

$$(h3)^+ \qquad \mathfrak{h}_3^+ \equiv D\mathfrak{H}_1^+ .$$

Remark. Some of the classes $\mathfrak{F}_i$, $\mathfrak{f}_i$, and $\mathfrak{h}_i$ have been introduced by a variety of different authors; some seem to be new.

(1) In the investigation of the bifurcation for the nonlinear Dirichlet boundary value problem (20), the class $\mathfrak{h}_1^+$ was introduced by Chafee [5] and Hale [6] and the class $\mathfrak{h}_2^\pm$ by Opial [11];

(2) For all functions in the class $\mathfrak{h}_1^+$ and C^2-functions in the class $\mathfrak{h}_3^+$, necessarily $f''(0) = 0$ holds. This rather severe restriction does not hold for typical asymmetric systems as (20) with $f(u) = e^u - 1$ below;

(3) The functional f_1 was introduced by Schaaf already in [14];

(4) The functional F_1 is used by Chicone [3] in the form $F_1 = f^4\,(F\,f^{-2})''$;

(5) The functional F_4 is introduced by Chow and Wang [4].

THEOREM 1. *The following inclusions hold among the sets of functions from Definitions* 2 *and* 4:

$$\begin{gathered}
\mathfrak{f}_i^+ \subset D\mathfrak{F}_i^+ \qquad \textit{for } i = 1,2,3,4 \\
\mathfrak{f}_1^+ \subset \mathfrak{f}_3^+ \subset \mathfrak{f}_4^+ \supset \mathfrak{f}_2^+ \supset \mathfrak{f}_1^+ \\
\mathfrak{F}_1^+ \subset \mathfrak{F}_3^+ \subset \mathfrak{F}_4^+ \supset \mathfrak{F}_2^+ \supset \mathfrak{F}_1^+ \cap \mathfrak{M} \\
\mathfrak{F}_5^+ \subset \mathfrak{F}_4^+ \\
\mathfrak{h}_1^+ \subset \mathfrak{f}_4^+ \cap \{\, f \mid f''(0) = 0 \,\} \\
\mathfrak{H}_1^+ \subset \mathfrak{F}_4^+ \cap \{\, f \mid f''(0) = 0 \,\}.
\end{gathered}$$

Assume for system (5) *that both functions f and g are contained in $D\mathfrak{F}_4^+$ and $\lambda, \mu > 0$.*

Then the energy-period function $T = T(E, \lambda, \mu)$ is increasing in E and satisfies $\partial T/\partial E > 0$ for all $E > 0$. The mean energy for the canonical ensemble given by (18) *is bounded below by $\langle H\rangle_{\text{can}} > \beta^{-1}$ for all inverse temperatures $\beta > 0$. If the functions f and g are contained in $D\overline{\mathfrak{F}}_4^+$, then the corresponding inequalities including equality hold.*

Remarks. (1) One way to set up a system with constant periods is to look for the borderline case of a potential $F \in \mathfrak{F}$ and force $f = F'$ such that either one of the six assumptions $F_i(q) = 0$ for $i = 1,2,3,4$ or $f_1(q) = 0$ or $f_4(q) = 0$ holds identically for all $q \in \mathbb{R}$. Solving the corresponding ordinary differential equations yields an explicit expression for the function F. From all six assumptions we get

$$F(q) = 9f'^3(0)f''^{-2}(0)\left[1 - \tfrac{1}{3}f''(0)f'^{-1}(0)\,q - \sqrt{1 - \tfrac{2}{3}f''(0)f'^{-1}(0)\,q}\right] \tag{24a}$$

in case $f''(0) \neq 0$ and

$$F(q) = \tfrac{1}{2}f'(0)q^2 \tag{24b}$$

in case $f''(0) = 0$. In the limit $f''(0) \to 0$, $F(q)$ of formula (24a) approaches (24b).

(2) Of course, there exist other systems with a center that has constant periods of oscillation. Loud [8] solves this problem for quadratic systems. Obi [10] gives a criterion for constant periods for system (6).

(3) The first terms of the Taylor expansion of the energy-period function in terms of the variable E can be calculated using normal forms, methods from bifurcation theory or the Laplace transform technique introduced below. We get

$$T(0, \lambda, \mu) = 2\pi\,(\lambda\mu f'(0)g'(0))^{-1/2} \tag{25a}$$

$$\begin{aligned} T'(0, \lambda, \mu) &= \frac{T(0, \lambda, \mu)}{24\lambda\mu}\left[\frac{5f''^2(0) - 3f'(0)f'''(0)}{f'^3(0)} + \frac{5g''^2(0) - 3g'(0)g'''(0)}{g'^3(0)}\right] \\ &= \frac{T(0, \lambda, \mu)}{24\lambda\mu}\left[\frac{f_1(0)}{f'^3(0)} + \frac{g_1(0)}{g'^3(0)}\right], \end{aligned} \tag{25b}$$

which gives the motivation to consider the quantity f_1.

(4) As a simple specific example, we get the expansion of the energy-period function for *Duffing's equation*

$$\frac{d^2x}{dt^2} + x + \beta x = 0. \tag{26}$$

It is equivalent to a Hamiltonian system (6) with $f(q) = q + \beta q^3$. Since $f'(0) = g'(0) = 1$, $g_1(p) \equiv 0$ and $f_1(q) \equiv -18\beta$, (25a) and (25b) yield the expansion

$$T(E) = 2\pi - \frac{3\pi}{2}\beta E + o(E). \tag{27}$$

This was derived by Loud [8] with a different approach.

THEOREM 2. *For the sets of functions from Definitions* 3 *and* 4, *the following inclusions hold*:

$$\mathfrak{f}_i^- \subset D\mathfrak{F}_i^- \quad \textit{for } i = 1 \textit{ and } 4,$$

$$\mathfrak{F}_1^- \subset \mathfrak{F}_3^- \subset \mathfrak{F}_4^- \supset \mathfrak{F}_2^- .$$

Assume for system (5) *that both functions f and g are contained in $D\mathfrak{F}_4^-$ and $\lambda, \mu > 0$.*

Then the energy-period function $T = T(E, \lambda, \mu)$ is decreasing in E and satisfies $\partial T/\partial E < 0$ for all $E > 0$. The mean energy for the canonical ensemble given by (18) *is bounded above by $\langle H \rangle_{\text{can}} < \beta^{-1}$ for all inverse temperatures $\beta > 0$. If the functions f and g are contained in $D\overline{\mathfrak{F}}_4^-$, then the corresponding inequalities including equality hold.*

The following theorem, based on more restrictive assumptions, deals with the dependence of the energy-period function and other quantities on the parameters λ and μ. It generalizes the results for the Volterra system from Rothe [13]. The indefinite integral of the energy-period function is denoted by

$$A(E, \lambda, \mu) = \int_0^E T(\widetilde{E}, \lambda, \mu)\, d\widetilde{E}. \tag{28}$$

It is equal to the area enclosed by the orbit $(p, q)(t, E, \lambda, \mu)$ in the (p, q)-plane.

THEOREM 3. *Assume for system* (5) *that both functions* f *and* g *are contained in* $D\mathfrak{F}_2^+$ *and* $\lambda, \mu > 0$. *Then the energy-period function* $T = T(E, \lambda, \mu)$ *and its integral* $A(E, \lambda, \mu)$ *have the following qualitative properties*:

$$\frac{\partial T}{\partial E} > 0 \quad \text{for all } E > 0, \tag{29}$$

$$\frac{\partial}{\partial E}\left(\frac{E}{T}\frac{\partial T}{\partial E}\right) > 0 \quad \text{for all } E > 0. \tag{30}$$

The functions

$$(\log(E/\lambda), \log(E/\mu)) \mapsto \log\left[ET(E, \lambda, \mu)\right], \tag{31}$$

$$(\log(E/\lambda), \log(E/\mu)) \mapsto \log A(E, \lambda, \mu) \tag{32}$$

are well defined and strictly convex on their entire domain of definition $\mathbb{R}^2$. *For the canonical ensemble with any inverse temperatures* $\beta > 0$, *the mean energy* $\langle H\rangle_{\text{can}}$ *and the specific heat* $\beta^2\text{Var}_{\text{can}}H$ *are bounded below by*

$$\beta^2\text{Var}_{\text{can}}H > \beta\langle H\rangle_{\text{can}} > 1. \tag{33}$$

Assume, furthermore, that $f'(q) > 0$ *for all* $q \in \mathbb{R}$ *and* $g'(q) > 0$ *for all* $q \in \mathbb{R}$. *Then for all* $\lambda, \mu > 0$ *and* $E \geq 0$, *we get further conclusions*:

$$\frac{\partial T}{\partial E} < \frac{T}{E}. \tag{34}$$

The limit $\lim_{E\to\infty} T'(E, \lambda, \mu) = T'(\infty, \lambda, \mu)$ *exists and*

$$ET'(\infty, \lambda, \mu) < T(E, \lambda, \mu) < T(0, \lambda, \mu) + ET'(\infty, \lambda, \mu) + o(E, \lambda, \mu), \tag{35}$$

where $\lim_{E\to\infty} o(E, \lambda, \mu) = 0$.

The orbital averages of the two summands of the Hamiltonian

$$\langle F(q)\rangle_{H=E} \equiv \frac{1}{T}\int_0^T F(q(t, E, \lambda, \mu))\, dt, \tag{36}$$

$$\langle G(p)\rangle_{H=E} \equiv \frac{1}{T}\int_0^T G(p(t, E, \lambda, \mu))\, dt \tag{37}$$

have a ratio in the range $[\frac{1}{2}, 2]$.

Remarks.

(1) For the harmonic oscillator with $f(q) = q$ and $g(p) = p$, formula (33) holds with equality signs.

(2) In the examples of classical mechanics (36) and (37) are the mean potential and kinetic energy. Hence we have shown equipartition of energy up to the factor of 2.

(3) A result similar to (30) is given by Theorem 1.4.2 of Schaaf [15] for the Dirichlet boundary value problem (20). Schaaf considers the period T as a function of the independent variable $p = \sqrt{2E}$. Under the assumption that f is an A-B-functions

(which holds for $f \in f_1{}^+$), Schaaf proves $d^2T/dp^2 > 0$ for all p. Conversion to the independent variable E via $d^2T/dp^2 = 2ET''(E) + T'(E)$ yields

$$\text{(S)} \qquad \frac{ET''}{T'} > -\frac{1}{2},$$

whereas (30) from above is equivalent to

$$\text{(R)} \qquad \frac{ET''}{T'} > \frac{ET'}{T} - 1.$$

We want to compare (R) and (S). In general, neither one implies the other. Indeed, take (20) and let $f(u) = e^u - 1$. For $E \to 0$, the right-hand side of (R) tends to -1, and hence (S) is stronger than (R). For $E \to \infty$, (34) and (35) imply that the right-hand side of (R) tends to zero. Hence (R) is stronger than (S).

PROPOSITION 4. *The following inclusions hold among the sets of functions from Definition* 5:

$$\mathfrak{h}_1^+ \cup \mathfrak{h}_3^+ \subset \mathfrak{h}_2^+.$$

Proof. The inclusion $\mathfrak{h}_1^+ \subset \mathfrak{h}_2^+$ is straightforward to prove using $(uf' - f)' = uf''$. We give the detailed proof that $\mathfrak{h}_3^+ \subset \mathfrak{h}_2^+$. Let $f \in \mathfrak{h}_3^+$ be given. The function $g = \sqrt{F}$ has the first derivative $g' = f/(2\sqrt{F})$ and the second derivative $g'' = (2Ff' - f^2)/(4\sqrt{F}^3) < 0$, which is negative since $f \in \mathfrak{h}_2^+$. Hence

$$(38) \qquad (uf' - f) = 2ugg'' + 2g'(ug' - g).$$

From l'Hôpital's rule, we calculate

$$(39) \qquad \lim_{u\to 0}(ug' - g) = \lim_{u\to 0} \frac{uf - 2F}{2\sqrt{F}} = \lim_{u\to 0} \frac{uf' - f}{(f/\sqrt{F})} = 0,$$

and hence

$$(40) \qquad (ug' - g) = \int_0^u sg''(s)\,ds.$$

Finally, (38), (39), and (40) imply the result, since

$$(41) \qquad (uf' - f) = 2ug'' + 2g' \int_0^u sg''(s)\,ds < 0.$$

3. Proof of Theorems 1 and 2. We define the functions $q = q_f(x)$ and $p = q_g(y)$ such that in the (noncanonical) coordinates (x, y), the orbits of system (5) are transformed into ellipses. The relevant transformation $q = q_f(x)$ depends only on the function f. The corresponding transformation $p = q_g(y)$ depending on the function g defines the second coordinate y.

The function $x \mapsto q_f(x)$ is uniquely defined by the requirements

$$(42) \qquad F(q_f(x)) = \tfrac{1}{2}x^2 \quad \text{and} \quad xq_f(x) \geq 0 \quad \text{for all } x \in \mathbb{R}.$$

Obviously $(F0)$ implies that $q_f(x) > 0$ is equivalent to $x > 0$, $q_f(x) = 0$ is equivalent to $x = 0$, and $q_f(x) < 0$ is equivalent to $x < 0$.

Furthermore, we introduce the functions $t_f(x)$ and $s_f(x)$ defined by

$$t_f(x) = \frac{dq_f(x)}{dx} \quad \text{and} \quad s_f(x) = \frac{t_f(x) + t_f(-x)}{2}, \tag{43}$$

as well as the analogous functions $t_g(y)$ and $s_g(y)$.

Differentiating (42) by x and solving for the derivatives of the function t_f yields the identities

$$f\left(q_f(8x)\right) t_f(x) = x, \tag{44}$$

$$f^3\left(q_f(x)\right) \frac{d t_f(x)}{dx} = f^2\left(q(x)\right) - 2F\left(q(x)\right) f'\left(q(x)\right), \tag{45}$$

$$\frac{f^5\left(q_f(x)\right)}{x} \frac{d^2 t_f(x)}{dx^2} = F_1\left(q(x)\right). \tag{46}$$

Especially $(f0)$ and (44) imply $t_f(x) > 0$ for all $x \neq 0$. Hence the function $q = q_f(x)$ is strictly increasing and the new coordinates (x, y) are well defined. Indeed $f'(0) > 0$ and (58) below imply that $t_f(0) > 0$ holds, too.

LEMMA 1. *In the new coordinates, system* (5) *is transformed to*

$$\begin{aligned} \frac{dx}{dt} &= \frac{\mu}{t_f(x) t_g(y)} y, \\ \frac{dy}{dt} &= -\frac{\lambda}{t_f(x) t_g(y)} x. \end{aligned} \tag{47}$$

The Hamiltonian given by (4) is a constant of motion which we call *energy* and denote by E. Transformed to the (x, y)-coordinates it is

$$E = H(p, q) = \lambda F(q) + \mu G(p) = \lambda \tfrac{1}{2} x^2 + \mu \tfrac{1}{2} y^2, \tag{48}$$

from which we see that the orbits of (47) are ellipses. In the (x, y)-plane, we use polar coordinates (E, φ) defined by

$$x = \sqrt{\frac{2E}{\lambda}} \cos\varphi, \qquad y = \sqrt{\frac{2E}{\mu}} \sin\varphi. \tag{49}$$

The equations of motion (47) transform to

$$\frac{d\varphi}{dt} = -\frac{\sqrt{\lambda\mu}}{t_f\left(\sqrt{2E/\lambda} \cos\varphi\right) t_g\left(\sqrt{2E/\mu} \sin\varphi\right)}, \tag{50}$$

which yields the following lemma, going back to Waldvogel [17] and Schaaf [14].

LEMMA 2. *The energy-period function of system* (5) *is*

$$T(E, \lambda, \mu) = (\lambda\mu)^{-1/2} \int_0^{2\pi} t_f\left(\sqrt{\frac{2E}{\lambda}} \cos\varphi\right) t_g\left(\sqrt{\frac{2E}{\mu}} \sin\varphi\right) d\varphi. \tag{51}$$

Remark. Formula (51) contains rather detailed information about the temporal structure of the oscillations. We divide the (p, q)-plane into the four quadrants $\{ q >$

$0, p > 0\}$, $\{q > 0, p < 0\}$, $\{q < 0, p < 0\}$, and $\{q < 0, p > 0\}$ and denote by T^{++}, T^{+-}, T^{--}, and T^{-+}, the parts of the period for which the orbits is in the four quadrants. With the same integrand a as in (51), we get

$$T^{++}(E,\lambda,\mu) = \int_0^{\pi/2} a\,d\varphi, \qquad T^{-+}(E,\lambda,\mu) = \int_{\pi/2}^{\pi} a\,d\varphi, \tag{52}$$

$$T^{--}(E,\lambda,\mu) = \int_{\pi}^{3\pi/2} a\,d\varphi, \qquad T^{+-}(E,\lambda,\mu) = \int_{3\pi/2}^{2\pi} a\,d\varphi. \tag{53}$$

For system (5) with nonsymmetric force f or g—as for example the Volterra system (1)—it may well happen that the energy-period function $T(E)$ is monotone increasing although some of the functions T^{++}, T^{+-}, T^{-+}, or T^{--} are not.

In (43) we have already introduced the function $s_f(x)$ in terms of which (51) can be rewritten as

$$T(E,\lambda,\mu) = 4\,(\lambda\mu)^{-1/2} \int_0^{\pi/2} s_f\left(\sqrt{\frac{2E}{\lambda}}\cos\varphi\right) s_g\left(\sqrt{\frac{2E}{\mu}}\sin\varphi\right) d\varphi. \tag{54}$$

Differentiating formula (54) by E yields the following.

LEMMA 3. *The derivative of the energy-period function is*

$$\frac{\partial T(E,\lambda,\mu)}{\partial E} = 4\,(\lambda\mu)^{-1/2} \int_0^{\pi/2} b\,d\varphi \quad \textit{with} \tag{55}$$

$$b = \frac{s_f{}'\left(\sqrt{\frac{2E}{\lambda}}\cos\varphi\right)}{\sqrt{\frac{2E}{\lambda}}\cos\varphi}\, s_g\left(\sqrt{\frac{2E}{\mu}}\sin\varphi\right)\cos^2\varphi + s_f\left(\sqrt{\frac{2E}{\lambda}}\cos\varphi\right) \frac{s_g'\left(\sqrt{\frac{2E}{\mu}}\sin\varphi\right)}{\sqrt{\frac{2E}{\mu}}\sin\varphi}\,\sin^2\varphi.$$

Formula (55) is the main tool of our proof. To begin with, we consider the special case $E = 0$ for which formulas (54) and (55) imply

$$T(0,\lambda,\mu) = 2\pi\,(\lambda\mu)^{-1/2}\, s_f(0)s_g(0), \tag{56}$$

$$\frac{\partial T(0,\lambda,\mu)}{\partial E} = \pi\,(\lambda\mu)^{-1/2}\left[s_f{}''(0)s_g(0) + s_f(0)s_g''(0)\right], \tag{57}$$

where we have used l'Hôpital's rule to get $\lim_{x\to 0+}(s_f(x))/x = s_f{}''(0)$. To express $s_f(0), s_f{}''(0)$ and $t_f(0), t_f{}'(0), t_f{}''(0)$ in terms of the function f and its derivatives, we solve formulas (44), (45), and (46) by the latter quantities and apply l'Hôpital's rule several times. For example,

$$t_f(0) = \lim_{x\to 0+} t_f(x) = \lim_{x\to 0+} \frac{x}{f\left(q_f(x)\right)}.$$

For this $\frac{0}{0}$-limit, l'Hôpital's rule gives

$$= \lim_{x\to 0+} \frac{1}{f'(q_f(x))\,q_f{}'(x)} = \frac{1}{f'(0)q_f{}'(0)}.$$

Since $t_f = q_f{}'$ by definition, we can solve for $t_f(0)$. Similar, but even longer calculations yield the derivatives $t_f{}'(0)$ and $t_f{}''(0)$. Finally, we end up with

$$t_f(0) = \frac{1}{\sqrt{f'(0)}}, \tag{58}$$

$$t_f{}'(0) = -\frac{f''(0)}{3f'^2(0)}, \tag{59}$$

$$t_f{}''(0) = \frac{5f''^2(0) - 3f'(0)f'''(0)}{12f'^{7/2}(0)} = \frac{f_1(0)}{12f'^{7/2}(0)}. \tag{60}$$

Plugging (58) and (60) into the formulas (56) for $T(0,\lambda,\mu)$ and (57) for $\partial T(0,\lambda,\mu)/\partial E$ yields (25a) and (25b).

We come to the main part of the proof based on (55) for $E > 0$. The monotonicity results all rely on the property

$$s_f{}'(x) > 0 \quad \text{for all} \quad x > 0, \tag{$(s)^+$}$$

which we shall conclude from the primary assumptions of Theorem 1. Obviously, if $(s)^+$ holds for both functions f and g, then the energy-period function $T = T(E,\lambda,\mu)$ satisfies $(\partial T/\partial E) > 0$ for all $E > 0$, and hence is increasing in E. The further discussion involves the following assumptions for the functions t_f or t_g:

$$t_f(x) - xt_f{}'(x) > 0 \quad \text{for all } x \in \mathbb{R}, \tag{$(t0)$}$$

$$t_f{}''(x) > 0 \quad \text{for all } x \neq 0, \tag{$(t1)^+$}$$

$$x\frac{d}{dx}\frac{xt_f{}'(x) - xt_f{}'(0)}{t_f(x) - xt_f{}'(0)} > 0 \quad \text{for all } x \neq 0, \tag{$(t2)^+$}$$

$$t_f(x) - xt_f{}'(x) < t_f(0) \quad \text{for all } x \neq 0, \tag{$(t3)^+$}$$

$$xt_f{}'(x) - xt_f{}'(0) > 0 \quad \text{for all } x \neq 0, \tag{$(t4)^+$}$$

$$\begin{gathered} t_f(x) - xt_f{}'(x) > 0 \quad \text{and} \quad x \neq 0 \quad \text{imply} \\ 3x(t_f'(x) - t_f'(0))\,(t_f(x) - xt_f{}'(x)) + x^2 t_f(x)t_f{}''(x) > 0. \end{gathered} \tag{$(t5)^+$}$$

Assumptions $(t1)^-$, $(t2)^-$, $(t3)^-$, and $(t4)^-$ are defined by the reversed inequalities, whereas $(si)^+$ and $(si)^-$ with $i = 0, 1, 2, 3, 4, 5$ denote the corresponding assumptions for the function s_f.

LEMMA 4. *Take a function $f \in \mathfrak{f}$. Then the following implications hold:*

$$(t1)^+ \implies (t3)^+ \implies (t4)^+ \implies (s)^+, \tag{61}$$

$$(t2)^+ \implies (t4)^+ \implies (s)^+, \tag{62}$$

$$(t1)^+ \quad \text{and} \quad (t0) \implies (t2)^+. \tag{63}$$

Proof of Lemma 4. We begin proving (61). The implication $(t1)^+ \Rightarrow (t3)^+$ is easy to get from $(t_f(x) - xt_f(x)')' = -xt_f(x)''$.

To prove $(t3)^+ \Rightarrow (t4)^+$, we use $x^2\ [(t_f(x) - t_f(0))/x]\ ' = t_f(0) - [t_f(x) - xt_f(x)']$. Indeed, $(t3)^+$ implies that the function $x \in (0,\infty) \mapsto (t_f(x) - t_f(0))/x$ is strictly increasing, and hence $t_f{}'(0) = \lim_{y\to 0}[(t_f(y) - t_f(0))/y] < [(t_f(x) - t_f(0))/x]$ for all $x > 0$. Together with a similar reasoning for $x < 0$, this yields $(t4)^+$.

The final implication $(t4)^+ \Rightarrow (s)^+$ arises by adding $(t4)^+$ to the corresponding inequality with x replaced by $-x$. Thus (61) is proved. We omit the more straightforward proof of (62).

To show (63), we assume that both $(t1)^+$and $(t0)$ hold. We define $\bar t_f(x) = t_f(x) - xt_f(0)$ and use the identity

$$x\bar t_f^2 \frac{d}{dx}\left(\frac{x\bar t_f{}'}{\bar t_f}\right) = x^2\,\bar t_f\bar t_f{}'' + x\bar t_f{}'\left(\bar t_f - x\bar t_f{}'\right). \tag{64}$$

We show that the right-hand side of (64) is positive for all $x > 0$. Indeed, $\bar t_f{}'' = t_f{}'' > 0$ by assumption $(t1)^+$, and hence $\bar t_f(x) > t_f(0)$ for all $x \neq 0$. Further on, $\bar t_f - x\bar t_f{}' = t_f - xt_f{}' > 0$ by assumption $(t0)$ and $x\bar t_f{}'(x) > 0$ for all $x \neq 0$ because of $(t4)^+$. Thus the right-hand side of (64) is positive for all $x > 0$, and hence the identity (64) implies $(t2)^+$ with $x > 0$. A similar reasoning applies to the case $x < 0$. □

LEMMA 5. *Assume* $f \in \mathfrak{n}$. *Then* $(t5)^+$ *implies* $(t4)^+$.

Proof. Since $f \in \mathfrak{n}$, there exist $q_1 \in [-\infty, 0)$ and $q_2 \in (0,\infty]$ such that

$$\begin{aligned} &f'(q) > 0 \quad \text{for all } q \in (q_1, q_2),\\ &f'(q) < 0 \quad \text{for all } q \in \mathbb{R} \setminus [q_1, q_2]. \end{aligned} \tag{n}$$

Let $x_1 \in [-\infty, 0)$ and $x_2 \in (0, \infty]$ correspond to q_1 and q_2 from assumption (n) via $q_f(x_i) = q_i$ for $i = 1, 2$. In the following argument we distinguish the cases

(a) $x \in [x_1, 0) \cup (0, x_2]$,

(b) $x \in (-\infty, x_1) \cup (x_2, \infty)$.

For $x \in [x_1, 0) \cup (0, x_2]$, the assumption $f \in \mathfrak{n}$ yields

$$\bar t_f - x\bar t_f{}' = t_t - xt_f{}' = (2F)^{3/2} f^{-3} f' > 0;$$

hence

$$\frac{1}{x}\frac{d}{dx}\left(\frac{x^3\,\bar t_f{}'}{\bar t_f - x\bar t_f{}'}\right) = \frac{x^2\,\bar t_f\bar t_f{}'' + 3x\bar t_f{}'\left(\bar t_f - x\bar t_f{}'\right)}{\left(\bar t_f - x\bar t_f{}'\right)^2} > 0, \tag{65}$$

and finally

$$x\bar t_f{}'(x) > 0 \quad \text{for all} \quad x \in [x_1, 0) \cup (0, x_2]. \tag{66}$$

Especially, $x_i\bar t_f{}'(x_i) = \bar t_f(x_i) > \bar t_f(0) > 0$ for $i = 1$ and $i = 2$. Now we turn to the case $x \in (-\infty, x_1) \cup (x_2, \infty)$. From (n) we get

$$x^2 \frac{d}{dx}\left(\frac{\bar t_f}{x}\right) = \bar t_f - x\bar t_f{}' = t_t - xt_f{}' = (2F)^{3/2} f^{-3} f' < 0, \tag{67}$$

implying

$$x\bar t_f{}'(x) > \bar t_f(x) > x\bar t_f(x_i)/x_i > 0 \quad \text{for all } x \in (-\infty, x_1) \cup (x_2, \infty). \tag{68}$$

Together, formula (66) and (68) imply

$$x\bar t_f{}'(x) = x\left(t_f{}'(x) - t_f{}'(0)\right) > 0$$

for all $x \neq 0$ proving $(t4)^+$. □

LEMMA 6. *Take a function $f \in \mathfrak{f}$. Then the following implications hold:*

$$(t1)^- \Longrightarrow (t3)^- \Longrightarrow (t4)^- \Longrightarrow (s)^-, \tag{69}$$

$$(t1)^- \Longrightarrow (s1)^- \Longrightarrow (s2)^- \Longrightarrow (s3)^- \Longrightarrow (s)^-. \tag{70}$$

We omit the proof, which is analogous to Lemma 4. Note that in each of the assumptions (ti) with $i = 1, 2, 3, 4$, equality for all $x \in \mathbb{R}$ holds if and only if we have the linear function $t_f(x) = t_f(0) + xt_f{}'(0)$.

The next step is to restate assumptions (ti) with $i = 1, 2, 3, 4$ in terms of the original function f and its primitive F.

LEMMA 7. *Take any $F \in \mathfrak{F}$, and let $f = F'$. Then*

(1) *$(t0)$ holds if and only if $f \in \mathfrak{m}$;*

(2) *For $i = 1, 2, 3, 4, 5$ assumption $(ti)^+$ holds if and only if $F \in \mathfrak{F}_i^+$ and the corresponding assumption including equality holds if and only if $F \in \overline{\mathfrak{F}}_i^{+}$;*

(3) *For $i = 1, 2, 3, 4$ assumption $(ti)^-$ holds if and only if $F \in \mathfrak{F}_i^-$ and the corresponding assumption including equality holds if and only if $F \in \overline{\mathfrak{F}}_i^{-}$.*

Proof. Recall $\bar{t}_f(x) = t_f(x) - xt_f(0)$, and let $q = q_f(x)$. Some lengthy calculations involving relations (44), (45), (46) and (58), (59), (60) yield the identities

$$t_f(x) - xt_f{}'(x) = (2F(q))^{3/2} f(q)^{-3} f'(q), \tag{$tF0$}$$

$$t_f{}''(x) = (2F(q))^{1/2} f(q)^{-5} F_1(q), \tag{$tF1$}$$

$$x\bar{t}_f^2 \frac{d}{dx}\left(\frac{x\bar{t}_f{}'}{\bar{t}_f}\right) = x^2 \bar{t}_f \bar{t}_f{}'' + x\bar{t}_f{}' \left(\bar{t}_f - x\bar{t}_f{}'\right) = 8F^2(q) f^{-6}(q) F_2(q), \tag{$tF2$}$$

$$t_f(0) - t_f(x) + xt_f{}'(x) = F_3(q), \tag{$tF3$}$$

$$x\bar{t}_f{}'(x) = xt_f{}'(x) - xt_f{}'(0) = \tfrac{1}{3} f'^2(0) f^3(q) F_4(q), \tag{$tF4$}$$

from which the lemma can be read off. □

LEMMA 8. *Let $F \in \mathfrak{F}$, $f = F'$, and $i = 1, 2, 3, 4$. Under that general assumption, $f \in \mathfrak{f}_i^+$ implies $(ti)^+$, and hence $F \in \mathfrak{F}_i^+$.*

Proof. We need the identities

$$\frac{d}{dq}\left(\frac{F_1(q)}{f'^{\,5/3}(q)}\right) = \frac{2F(q)f(q)}{3f'^{\,8/3}(q)} f_1(q), \tag{71}$$

$$\frac{d}{dq}\left(\frac{F_4(q)}{f'(q)}\right) = \frac{f^2(q)}{f'^2(q)} f_4(q). \tag{72}$$

To begin with, we give the proof for the case $i = 4$. Assume that $f \in \mathfrak{f}_4^+$. By Definition 2, there exist $q_1 \in [-\infty, 0)$ and $q_2 \in (0, \infty]$ such that

(a) $f'(q) > 0$ and $f_4(q) > 0$ for all $q \in (q_1, 0) \cup (0, q_2)$;

(b) $f'(q) < 0$ for all $q \in \mathbb{R} \setminus [q_1, q_2]$.

Let $x_1 \in [-\infty, 0)$ and $x_2 \in (0, \infty]$ correspond to q_1 and q_2 via $q_f(x_i) = q_i$ for $i = 1, 2$. At first, we consider item (a). From the identity (72) we get $F_4(q) > 0$ for all $q \in (q_1, 0) \cup (0, q_2)$, and hence identity $(tF4)$ implies

$$x\bar{t}_f{}'(x) > 0 \quad \text{for all} \quad x \in (x_1, 0) \cup (0, x_2). \tag{73}$$

Now consider item (b). From $(tF0)$ we conclude $\bar{t}_f(x) - x\bar{t}_f{}'(x) < 0$, and hence

$$x\bar{t}_f{}'(x) > \bar{t}_f(x) > 0 \quad \text{for all} \quad x \in (-\infty, x_1) \cup (x_2, \infty). \tag{74}$$

Since $x\bar{t}_f{}'(x) = xt_f{}'(x) - xt_f{}'(0)$, (73), (74) imply $(t4)^+$.

To give the proof for the case $i = 3$, assume that $f \in \mathfrak{f}_3^+$. By Definition 2, there exist $q_1 \in [-\infty, 0)$ and $q_2 \in (0, \infty]$ such that

(a) $f'(q) > 0$ and $f_1(q) > 0$ for all $q \in (q_1, 0) \cup (0, q_2)$;

(b) $f'(q) < 0$ for all $q \in \mathbb{R} \setminus [q_1, q_2]$.

At first, we consider item (a). From (60) we get $t_f{}''(0) > 0$. Since $F_1(0) = 0$, the identity (71)and the assumption $f_1(q) > 0$ imply $F_1(q) > 0$ for all $q \in (q_1, 0) \cup (0, q_2)$. Hence $(tF1)$ yields $t_f{}''(x) > 0$, and hence

$$t_f(x) - xt_f{}'(x) < t_f(0) \quad \text{for all} \quad x \in [x_1, 0) \cup (0, x_2]. \tag{75}$$

Now consider item (b). From $f'(q) < 0$ and $(tF0)$ we conclude that

$$t_f(x) - xt_f{}'(x) < 0 \quad \text{for all} \quad x \in (-\infty, x_1) \cup (x_2, \infty). \tag{76}$$

Finally, (75), (76) imply $(t3)^+$.

For the proof of case $i = 2$, assume $f \in \mathfrak{f}_2^+$. Clearly $\mathfrak{f}_2^+ \subset \mathfrak{f}_3^+$, and hence (73)–(76) and $(t3)^+$, $(t4)^+$ all hold. Furthermore, by Definition 2, there exist $q_1 \in [-\infty, 0)$ and $q_2 \in (0, \infty]$ such that

(a) $f'(q) > 0$ and $f_1(q) > 0$ for all $q \in (q_1, q_2)$;

(b) $f'(q) < 0$ and $f_4(q) > 0$ for all $q \in \mathbb{R} \setminus [q_1, q_2]$.

Consider item (a) first. From $f'(q) > 0$ and $(tF0)$ we conclude that $\bar{t}_f - x\bar{t}_f{}' = t_f(x) - xt_f{}'(x) > 0$. Hence (64), (75) imply

$$x\,\bar{t}_f^{-2} \frac{d}{dx}\left(\frac{x\bar{t}_f{}'}{\bar{t}_f}\right) = x^2\,\bar{t}_f\bar{t}_f{}'' + x\bar{t}_f{}'\left(\bar{t}_f - x\bar{t}_f{}'\right) > 0 \tag{77}$$

for all $x \in (x_1, 0) \cup (0, x_2)$. Now turn to item (b). Because of the identity

$$f^2(q)f'^{-2}(q)\,f_4(q) = 3f'^2(0)x\,t_f^{-1}(x)\left(t_f(x) - xt_f{}'(x)\right)^{-2}\ B$$
$$\text{with}\ \ B = x\,\bar{t}_f^{-2}\frac{d}{dx}\left(\frac{x\bar{t}_f{}'}{\bar{t}_f}\right) + 2x\bar{t}_f{}'\left(\bar{t}_f - x\bar{t}_f{}'\right),$$

we can conclude from $f_4(q) > 0$ that $B > 0$. Now $F'(q) < 0$ implies $\bar{t}_f - x\bar{t}_f < 0$, and hence

$$x\,\bar{t}_f^{-2}\frac{d}{dx}\left(\frac{x\bar{t}_f{}'}{\bar{t}_f}\right) > 0 \quad \text{for all } x \in (-\infty, x_1) \cup (x_2, \infty). \tag{78}$$

Together, (77) and (78) imply $(t2)^+$.

Finally, to consider the case $i = 1$, we assume $f \in \mathfrak{f}_1^+$. Note that $f \in \mathfrak{f}_2^+$ holds and $q_1 = -\infty, q_2 = +\infty$. Hence we argue as in case $i = 2$ with $x_1 = -\infty, x_2 = +\infty$. □

LEMMA 9. *Again we need the general assumption $F \in \mathfrak{F}$, $f = F'$. Then $f \in \mathfrak{f}_i^-$ implies $(ti)^-$ for $i = 1$ and $i = 4$.*

We skip the proof, which is similar to that of Lemma 8. Lemmas 4–9 prove the inclusions of the various function classes in Theorem 1 and 2. The monotonicity of the energy-period function follows from Lemma 3 and the monotonicity of the functions s_f and s_g. To prove the estimate of $\langle H \rangle_{\text{can}}$, we use (18) and (19) and a partial integration to get

$$\beta\,\langle H \rangle_{\text{can}} - 1 = \frac{1}{Z}\int_0^\infty e^{-\beta E} E \frac{dT}{dE}\, dE.$$

This completes the proofs of Theorems 1 and 2.

4. Proof of Theorem 3. Recall formula (43), defining $s_f(x) = \frac{1}{2}(t_f(x) + t_f(-x))$. Let $E = \frac{1}{2}x^2$, and, as in Rothe [12], introduce the function

$$\tau_f(E) = \sqrt{\tfrac{2}{E}}\, s_f(\sqrt{2E}). \tag{79}$$

The *convolution* of two functions $E \in [0, \infty) \mapsto u(E)$ and $E \in [0, \infty) \mapsto v(E)$ is defined as

$$(u * v)(E) = \int_0^E u(a)v(E-a)\, da. \tag{80}$$

We use the shorthand notation $\{u(E/\mu)\}$ for the function $E \mapsto u(E/\mu)$, where $\mu \in \mathbb{R}$ is a parameter. These definitions allow us to rewrite the formula (51) for the period as

(81)
$$T(E,\lambda,\mu) = (\lambda\mu)^{-1} \int_0^E \tau_f\left(\frac{a}{\lambda}\right) \tau_g\left(\frac{E-a}{\mu}\right) da = (\lambda\mu)^{-1} \left\{\tau_f\left(\frac{E}{\lambda}\right)\right\} * \left\{\tau_g\left(\frac{E}{\mu}\right)\right\}.$$

As formula (19) from the introduction shows, the canonical partition function $Z(\beta, \lambda, \mu)$ is the Laplace transform of the energy-period function $T(E, \lambda, \mu)$. For Laplace correspondences like (19) we introduce the symbolic notation $T(E, \lambda, \mu) \;\bullet\!\!-\!\!\circ\; Z(\beta, \lambda, \mu)$. Here β denotes the independent variable of the Laplace transformed function. It has the physical meaning of an inverse absolute temperature. For a given function $F \in \mathfrak{f}$ and $f = F'$ in system (5), we define

$$z_f(\beta) = \int_0^\infty \exp\left(-\beta F(q)\right) dq. \tag{82}$$

We assemble some useful Laplace correspondences.

Lemma 10. *The following functions are Laplace transforms of each other.*

$$\tau_f(E) \;\bullet\!\!-\!\!\circ\; z_f(\beta), \tag{83}$$

$$\sqrt{\tfrac{2}{E}} \;\bullet\!\!-\!\!\circ\; \sqrt{2\pi/\beta}, \tag{84}$$

$$\tau_f\left(\frac{E}{\lambda}\right) \;\bullet\!\!-\!\!\circ\; \lambda z_f(\lambda\beta), \tag{85}$$

$$T(E, \lambda, \mu) \;\bullet\!\!-\!\!\circ\; Z(\beta, \lambda, \mu). \tag{86}$$

Proof. Substituting $E = F(p)$ into (82) yields

$$z_f(\beta) = \int_0^\infty \exp\left(-\beta E\right) \tilde{\tau}_f(E)\, dE, \tag{87}$$

with $\tilde{\tau}_f(E) = \sum\{\,|F'(q)|^{-1} \mid F(q) = E\,\} = \tau_f(E)$. This proves (83) from which (85) follows. The correspondence (84) is elementary. Alternatively, it follows from (83) for the harmonic oscillator case $f(q) = q$ and $F(q) = \frac{1}{2}q^2$. Formula (86) was already proved in the introduction. □

Since the Laplace transformation takes convolutions into products, applying (85) to both functions f and g yields

$$\left\{\tau_f\left(E/\lambda\right)\right\} * \left\{\tau_g\left(E/\mu\right)\right\} \;\bullet\!\!-\!\!\circ\; \lambda\mu z_f(\lambda\beta) z_g(\mu\beta). \tag{88}$$

Because the Hamiltonian is assumed to be the sum (4), the canonical partition function defined by (17) factors into a product and we get

$$Z(\beta,\lambda,\mu)=z_f(\lambda\beta)z_g(\mu\beta). \tag{89}$$

From (86)–(89), we get an independent proof of the energy-period formula (81).

After these more general considerations, we are ready to begin the proof. In Theorem 3, we make the assumption that $f\in D\mathfrak{F}_2^+$ and $g\in D\mathfrak{F}_2^+$. Consider the function f. By Lemma 7, the inequality $(t2)^+$ holds, which asserts that both functions $\log x\mapsto\log\left(t_f(x)-xt_f{}'(0)\right)$ and $\log x\mapsto\log\left(t_f(-x)+xt_f{}'(0)\right)$ are strictly convex on their whole domain $\mathbb{R}\ni x$. From Rothe [13], we need the following.

LEMMA 11. *Let $f=f(x)$ and $g=g(x)$ be smooth positive functions defined for all $x\in(0,\infty)$. If the two functions $\log x\mapsto\log f(x)$ and $\log x\mapsto\log g(x)$ are both strictly convex on their domain $\mathbb{R}$, then the following functions are strictly convex on their entire domains*:

$$\log x\in\mathbb{R}\mapsto\log\left(f(x)+g(x)\right), \tag{90}$$

$$\log x\in\mathbb{R}\mapsto\log\left(f(x)g(x)\right), \tag{91}$$

$$(\log x,\log y)\in\mathbb{R}^2\mapsto\log\left[\int_0^1 f(sx)g((1-s)y)\,ds\right], \tag{92}$$

$$(\log(E/\mu),\log(E/\lambda))\in\mathbb{R}^2\mapsto\log\left[(\lambda\mu)^{-1/2}\{f(E/\lambda)\}*\{g(E/\mu)\}\right]. \tag{93}$$

Item (93)—not contained in Rothe [13]—is a consequence of (92). Note that the factor $(\lambda\mu)^{-1/2}$ makes the function in formula (93) well defined. By means of (90) in Lemma 11, we conclude that the function $\log x\in\mathbb{R}\mapsto\log s_f(x)$ is strictly convex. Now it is straightforward from (79) that both functions

$$\log\left(\frac{E}{\lambda}\right)\mapsto\log\tau_f\left(\frac{E}{\lambda}\right)\quad\text{and}\quad\log\left(\frac{E}{\mu}\right)\mapsto\log\tau_g\left(\frac{E}{\mu}\right) \tag{94}$$

are strictly convex, and finally (92) in Lemma 11 implies that the function

$$\left(\log\left(\frac{E}{\mu}\right),\log\left(\frac{E}{\lambda}\right)\right)\mapsto\log[(\lambda\mu)^{-1/2}T(E,\lambda,\mu)]$$

is strictly convex on its domain $\mathbb{R}^2$, which proves (31).

Next we prove estimate (32) for the area enclosed by the orbits of system (5). Define

$$T_f(E)=\int_0^\infty\sqrt{\tfrac{2}{a}}\,\tau_f(E-a)\,da=\left\{\sqrt{\tfrac{2}{E}}\right\}*\{\tau_f(E)\}, \tag{95}$$

and similarly T_g corresponding to g. Indeed, T_f is the energy-period function of the system

$$\begin{aligned}\frac{dq}{dt}&=p,\\ \frac{dp}{dt}&=-f(q),\end{aligned} \tag{96a}$$

and T_g is the energy-period function of the system

$$\begin{aligned} \frac{dq}{dt} &= g(p), \\ \frac{dp}{dt} &= -q. \end{aligned} \tag{96b}$$

A formula for the enclosed area A as a convolution can be found by means of Laplace transforms. From $A(E,\lambda,\mu) = \int_0^\infty T(E,\lambda,\mu)\,dE$ and elementary properties of the Laplace transformation, we get the correspondence

$$A(E,\lambda,\mu) \;\bullet\!\!-\!\!\circ\; \beta^{-1}Z(\beta,\lambda,\mu). \tag{97}$$

On the other hand, taking the convolution of (84) and (85) yields $\{\sqrt{2/E}\}*\{\tau_f(E)\} = T_f(E) \;\bullet\!\!-\!\!\circ\; \sqrt{2\pi/\beta}\,z_f(\beta)$, and hence

$$T_f\left(\frac{E}{\lambda}\right) \;\bullet\!\!-\!\!\circ\; \lambda\sqrt{2\pi/\lambda\beta}\,z_f(\lambda\beta) \quad\text{and}\quad T_g\left(\frac{E}{\mu}\right) \;\bullet\!\!-\!\!\circ\; \mu\sqrt{2\pi/\mu\beta}\,z_g(\mu\beta).$$

Taking the convolution of these two formulas yields

$$\begin{aligned} \{T_f\,(E/\lambda)\} * \{T_g\,(E/\mu)\} \;\bullet\!\!-\!\!\circ\; & 2\pi\lambda\mu\,\frac{z_f(\lambda\beta)}{\sqrt{\lambda\beta}}\,\frac{z_g(\mu\beta)}{\sqrt{\mu\beta}} \\ &= 2\pi\sqrt{\lambda\mu}\,\beta^{-1}Z(\beta,\lambda,\mu). \end{aligned} \tag{98}$$

Now comparison of (97) and (98) yields

$$A(E,\lambda,\mu) = \frac{1}{2\pi\sqrt{\lambda\mu}}\left\{T_f\left(\frac{E}{\lambda}\right)\right\} * \left\{T_g\left(\frac{E}{\mu}\right)\right\}. \tag{99}$$

Applying formula (31), which we have already proved above to the systems (96a) and (96b), we conclude that both functions

$$\log\left(\frac{E}{\lambda}\right) \mapsto \log T_f\left(\frac{E}{\lambda}\right) \quad\text{and}\quad \log\left(\frac{E}{\mu}\right) \mapsto \log T_g\left(\frac{E}{\mu}\right)$$

are strictly convex. Now (93) from Lemma 11 yields convexity for the convolution. Because of (99), we get that the function

$$\left(\log\left(\frac{E}{\mu}\right), \log\left(\frac{E}{\lambda}\right)\right) \mapsto \log A(E,\lambda,\mu)$$

is strictly convex on its domain $\mathbb{R}^2$, which proves (32).

The results about the canonical ensemble rely on the following lemma proved in [13, p. 676].

LEMMA 12. *Assume that the function* $\log E \in \mathbb{R} \mapsto \log\tau(E)$ *is strictly convex on* $\mathbb{R}$. *Taking the Laplace transform* $z(\beta) = \int_0^\infty e^{-\beta\tau(E)}\,dE$, *we get the function* $\log\beta \mapsto \log z(\beta)$, *which is strictly convex on* $\mathbb{R}$.

By this Lemma, the functions $\log\beta \mapsto \log z_f(\beta)$ and $\log\beta \mapsto \log z_g(\beta)$ are strictly convex. Since $Z(\beta,\lambda,\mu) = z_f(\lambda\beta)z_g(\mu\beta)$, by Lemma 12, the function $\log\beta \mapsto \log Z(\beta,\lambda,\mu)$ is strictly convex, too. Using (18) we conclude

$$0 < \beta\frac{\partial}{\partial\beta}\left(\beta\frac{\partial}{\partial\beta}\log Z\right) = \beta^2\mathrm{Var}_{\mathrm{can}}H - \beta\,\langle H\rangle_{\mathrm{can}}$$

for all $\beta > 0$. Since $\lim_{\beta\to 0}\beta\,\langle H\rangle_{\text{can}} = 1$, we get assertion (13).

We come to the second part of Theorem 3, relying on the additional assumption $f'(q) > 0$ and $g'(p) > 0$ for all $p, q \in \mathbb{R}$ as well as $f, g \in D\mathfrak{F}_2^+$ as before. Let s_f be the function defined by (43). Under the assumptions just stated, the function $x \in (0,\infty) \mapsto x s_f{}'(x)/s_f(x)$ is increasing and bounded above by the constant 1 because of $(tF0)$. Hence the function $x \in (0,\infty) \mapsto s_f(x)/x$ is positive and decreasing and finally $\lim_{x\to\infty} s_f{}'(x) = \lim_{x\to\infty} s_f\,(x)/x = s'_\infty$ exists. For all $\delta > 0$ there exists K_δ such that

$$xs'_\infty < s_f(x) < (1+\delta)xs'_\infty + K_\delta \quad \text{for all } x > 0. \tag{100}$$

Inserting this estimate into the basic period formula (54) yields the estimate

$$\begin{aligned}\left(\frac{4E^2}{\lambda\mu}\right) s'_{f\infty}s'_{g\infty} &< ET(E,\lambda,\mu)\\ &<(1+\delta)\left(\frac{4E^2}{\lambda\mu}\right) s'_{f\infty}s'_{g\infty}\\ &\quad + K'_\delta\left(1+\left(\frac{E}{\lambda}\right)^{3/2}+\left(\frac{E}{\mu}\right)^{3/2}\right).\end{aligned} \tag{101}$$

For the proof of the last claim in Theorem 3 we need the following.

LEMMA 13. *The orbital averages of F and G are*

$$\begin{aligned}\langle F(q)\rangle_{H=E} &\equiv \frac{1}{T}\int_0^T F(q(t,E))\,dt = -\frac{1}{T}\frac{\partial A(E,\lambda,\mu)}{\partial\lambda},\\ \langle G(p)\rangle_{H=E} &\equiv \frac{1}{T}\int_0^T G(p(t,E))\,dt = -\frac{1}{T}\frac{\partial A(E,\lambda,\mu)}{\partial\mu},\end{aligned} \tag{103}$$

or, in terms of the phase angle,

$$\begin{aligned}\langle F(q)\rangle_{H=E} &= \frac{4E}{T\lambda\sqrt{\lambda\mu}}\int_0^{\pi/2} s_f\left(\sqrt{\frac{2E}{\lambda}}\cos\varphi\right) s_g\left(\sqrt{\frac{2E}{\mu}}\sin\varphi\right)\cos^2\varphi\,d\varphi,\\ \langle G(p)\rangle_{H=E} &= \frac{4E}{T\mu\sqrt{\lambda\mu}}\int_0^{\pi/2} s_f\left(\sqrt{\frac{2E}{\lambda}}\cos\varphi\right) s_g\left(\sqrt{\frac{2E}{\mu}}\sin\varphi\right)\sin^2\varphi\,d\varphi.\end{aligned} \tag{104}$$

Proof. Differentiating the partition function $Z(\beta,\lambda,\mu)$ in (17) by the parameter λ and substituting the independent variables (t, E) yields

$$-\frac{1}{\beta}\frac{\partial Z(\beta,\lambda,\mu)}{\partial\lambda} = \int_0^\infty e^{-\beta E}\left(\int_0^{T(E,\lambda,\mu)} F(q(t,E,\lambda,\mu))dt\right)dE, \tag{105}$$

which is the Laplace correspondence $-T(E,\lambda,\mu)\,\langle F(q)\rangle_{H=E}$ •—○ $\beta^{-1}\partial Z/\partial\lambda$. On the other hand, differentiating the correspondence (97) by λ yields $\partial A(E,\lambda,\mu)/\partial\lambda$ •—○ $\beta^{-1}\partial Z/\partial\lambda$. Comparison of the last two formulas proves (103).

To show (104), we use (50) to substitute the phase angle φ into the definition (103) of $\langle F(q)\rangle$ and plug in $F(q) = \frac{1}{2}x^2 = (E/\lambda)\cos^2\varphi$, resulting from (42) and (49). □

To complete the proof of Theorem 3, we show (35) through (37). As further consequences of (100),

(106)
$$\left(\frac{2E^2}{\lambda\mu}\right) s'_{f\infty} s'_{g\infty} < X < (1+\delta)\left(\frac{2E^2}{\lambda\mu}\right) s'_{f\infty} s'_{g\infty} + K''_\delta\left(1+\left(\frac{E}{\lambda}\right)^{3/2}+\left(\frac{E}{\mu}\right)^{3/2}\right)$$

holds for all $(E/\lambda) > 0, (E/\mu) > 0$ uniformly—with the three choices $X = A(E,\lambda,\mu)$ as well as $X = \lambda ET\,\langle F(q)\rangle_{H=E}$ and $X = \mu ET\,\langle G(p)\rangle_{H=E}$. The first choice results from integrating (101) over the interval $[0,E]$; the two others use (104). Together, (101) and (106) imply

$$\lim_{E\to\infty}\frac{\lambda T}{A}\,\langle F(q)\rangle_{H=E} < 1 \quad\text{and}\quad \lim_{E\to\infty}\frac{\mu T}{A}\,\langle G(p)\rangle_{H=E} < 1. \tag{107}$$

On the other hand, elementary considerations show that

$$\lim_{E\to 0}\frac{\lambda T}{A}\,\langle F(q)\rangle_{H=E} = \frac{1}{2} \quad\text{and}\quad \lim_{E\to\infty}\frac{\mu T}{A}\,\langle G(p)\rangle_{H=E} = \frac{1}{2}. \tag{108}$$

Finally, we combine this information about the limits $E \to \infty$ and $E \to 0$ with the convexity from (32) and (103),

$$\frac{\partial}{\partial\lambda}\frac{\lambda}{A}\frac{\partial A}{\partial\lambda} = -\frac{\partial}{\partial\lambda}\frac{\lambda T}{A}\,\langle F(q)\rangle > 0 \quad\text{and}\quad \frac{\partial}{\partial\mu}\frac{\mu}{A}\frac{\partial A}{\partial\mu} = -\frac{\partial}{\partial\mu}\frac{\mu T}{A}\,\langle G(p)\rangle > 0,$$

to conclude that

$$\frac{\lambda T}{A}\,\langle F(q)\rangle_{H=E} \quad\text{and}\quad \frac{\mu T}{A}\,\langle G(p)\rangle_{H=E} \in (1/2,1) \quad\text{for all}\quad E,\lambda,\mu > 0, \tag{109}$$

proving (36) and (37) in Theorem 3.

A similar reasoning, combining convexity and limits for $E \to \infty$, applies to derivatives by E. Because of (30), the function $E \mapsto (E/T)(dT/dE)$ is increasing. Because of (101) we get $\lim_{E\to\infty}(E/T)(dT/dE) \leq 1$. Hence $\lim_{E\to\infty} T(E)/E = \lim_{E\to\infty} dT/dE = T'_\infty$ exists, and estimate (35) follows.

5. Examples.

(A) *The pendulum.* The frictionless pendulum is a Hamiltonian system (2) with

$$H(p,q) = \tfrac{1}{2}p^2 + 1 - \cos q. \tag{A.1}$$

In the (p,q) phase plane, there exist closed orbits surrounding the center at $(0,0)$ with $H(0,0) = 0$. We restrict the discussion to these oscillations and disregard the librations occuring at energies $H > 2$. The closed oscillatory orbits lie inside the bounded region $\mathcal{R}$ bounded by the two heteroclinic orbits connecting the saddle points S_1 at $(p,q) = (0,-\pi)$ with S_2 at $(p,q) = (0,\pi)$. At the saddle points and on the connecting heteroclinic orbits, the Hamiltonian assumes the value $H(p,q) = 2$. Hence the restriction to the region $\mathcal{R}$ implies $E \in [0,2)$ for the range of values of the Hamiltonian and $q \in (-\pi,\pi)$ for the phase angle in the potential $F(q) = 1 - \cos q$ and the force $f(q) = \sin q$.

Referring to Definition 1–4, we check that $f \in \mathfrak{n}$ but $f \notin \mathfrak{m}$, $f \in \mathfrak{h}_1^+$, and $f \in \mathfrak{f}_2^+$. (Hence even $f \in \mathfrak{f}_3^+$ and $f \in \mathfrak{f}_4^+$.) Therefore, Theorem 1 shows that the energy-period function $T = T(E)$ is strictly increasing. Since even $f \in \mathfrak{f}_2^+$ does hold, we apply Theorem 3 to get the stronger conclusion that

$$\frac{\partial}{\partial E}\left(\frac{E}{T}\frac{\partial T}{\partial E}\right) > 0 \quad \text{for all} \quad E \in [0,2).$$

Because of $f \notin \mathfrak{m}$, the second part of Theorem 3 leading to (34)–(37) does not apply. Indeed (34) and (35) do not hold for the pendulum. This is easy to see because the slow motion near the saddle points S_1 and S_2 implies $\lim_{E\to 2} T(E) = \infty$.

(B) *The "battle of sexes" model.* See, e.g., Maynard-Smith and Hofbauer [9] for the biological background of the "battle of sexes" model. Let the variables $u, v \in [0,1]$ give the parts of the male and female population with different mating behavior. Taking into account the effects of behavior on the numbers of offspring, Maynard-Smith derives the differential equations

$$\text{(B.1)} \qquad \begin{aligned} \frac{du}{dt} &= \lambda u\,(1-v)\left(1-\frac{v}{B}\right), \\ \frac{dv}{dt} &= -\mu v\,(1-u)\left(1-\frac{u}{A}\right) \end{aligned}$$

with parameters $\lambda, \mu > 0$ and $A, B \in [0,1)$. A transformation to a Hamiltonian system (2) is achieved by introducing the variables p and q via

$$\text{(B.2)} \qquad u = \frac{A}{A + (1-A)e^{-p}} \quad \text{and} \quad v = \frac{B}{B+(1-B)e^{-q}}.$$

(B.2) transforms (B.1) into the system (2)—or equivalently (5). The relevant Hamiltonian has the form (4) of a sum with separated variables. Indeed, $H(p,q) = \lambda F(q) + \mu G(p)$ with

(B.3)

$$F(q) = -q + A^{-1}\log\left[1 + A(e^q - 1)\right] \quad \text{and} \quad G(p) = -p + B^{-1}\log\left[1 + B(e^p-1)\right].$$

For the function $f(q) = F'(q) = -(A-1)(e^q-1)/[1+A(e^q-1)]$, we have to check that $f_1 = 5f''^2 - 3f'f''' \geq 0$. This leads to a rather lengthy calculation, unless we use the Schwarzian derivative

$$\mathcal{S}f \equiv \frac{f'''}{f'} - \frac{3}{2}\left(\frac{f''}{f'}\right)^2$$

for a shortcut. Indeed, since $f(q) = -(A-1)U/[1+AU] \equiv T(U)$ with $U(q) = e^q - 1$, the function f is the composition $f = T \circ U$, where T is a linear fractional transformation, for which the Schwarzian derivative vanishes. Hence we get for the Schwarzian of the composition,

$$\mathcal{S}f = \mathcal{S}(T\circ U) = [\mathcal{S}T \circ U]U'^2 + \mathcal{S}U = \mathcal{S}U = -\tfrac{1}{2},$$

and finally $f_1 = (f''^2/2) - 3f'^2\mathcal{S}f = (f''^2 + 3f'^2)/2 > 0$ for all $q \neq 0$. (Indeed, $f_1 = 2A^{-2}(1-A)^{-2}W^2(1+W+W^2)(1+W)^{-6} > 0$, where $W = A(1-A)^{-1}e^q$.) Since $f, g \in \mathfrak{m}$, we have confirmed that $f, g \in \mathfrak{f}_1^+$. Thus all parts of Theorem 3 do apply. As a final remark, note that the limiting case $A = B = 0$ leads back to the original Volterra model (1).

(C) *The power law.* Take the Hamiltonian

$$H(p,q) = \lambda \frac{|q|^\sigma}{\sigma} + \mu \frac{|p|^\rho}{\rho} \tag{C.1}$$

of the form (4) with the power law potentials $F(q) = |q|^\sigma/\sigma$ and $G(p) = |p|^\rho/\rho$. This example shows the advantage gained from the partition function (17) to achieve an easy calculation of the energy-period function. At first we use (82) to get

$$z_f(\gamma) = \int_0^\infty \exp\left(-\gamma F(q)\right)\, dq \; = \; 2\sigma^{-1+1/\sigma}\Gamma(\sigma^{-1})\gamma^{-1/\sigma},$$

where Γ is the Euler gamma function. Now formula (88) yields

$$\begin{aligned} Z(\beta,\lambda,\mu) &= z_f(\lambda\beta)z_g(\mu\beta) \\ &= 4\sigma^{-1+1/\sigma}\rho^{-1+1/\rho}\Gamma(\sigma^{-1})\Gamma(\rho^{-1})\;(\lambda\beta)^{-1/\sigma}\,(\mu\beta)^{-1/\rho}. \end{aligned} \tag{C.2}$$

Now we use the Laplace correspondence $T(E,\lambda,\mu)$ •—∘ $Z(\beta,\lambda,\mu)$ from (19) to transform (C.2) back and get

$$ET(E,\lambda,\mu) = 4\sigma^{-1+1/\sigma}\rho^{-1+1/\rho}\Gamma(\sigma^{-1})\Gamma(\rho^{-1})\Gamma(\sigma^{-1}+\rho^{-1})\left(\frac{E}{\lambda}\right)^{1/\sigma}\left(\frac{E}{\mu}\right)^{1/\rho}.$$

Integration over the interval $[0,E]$ and formula (103) yields

$$A = \frac{ET}{\sigma^{-1}+\rho^{-1}} \quad \text{and} \quad \lambda\sigma\left\langle F(q)\right\rangle_{H=E} = \mu\rho\left\langle G(q)\right\rangle_{H=E} = \frac{E}{\sigma^{-1}+\rho^{-1}}.$$

In the case $\sigma,\rho \in (1,2)$, Theorem 1 and the first part of Theorem 3 apply. In the case $\sigma,\rho \in (2,\infty)$, Theorem 2 applies.

(D) *An unsymmetric system with constant periods and its perturbation.* Constant periods arise for the potential (24a). For simplicity, let $f'(0) = 1$, $f''(0) = -3$, and take the kinetic energy $G(p) = \frac{1}{2}p^2$ with $\mu = 1$. We get the Hamiltonian

$$H_0(p,q) = \tfrac{1}{2}p^2 + \lambda\left(1 + q - \sqrt{1+2q}\right). \tag{D.1}$$

System (2) with the Hamiltonian (D.1) leads to oscillations in the range $q \in [-\frac{1}{2}, \frac{3}{2}]$, since the potential $F(q) = \lambda(1 + q - \sqrt{1+2q})$ is no longer well defined for $q < -\frac{1}{2}$ and $F(-\frac{1}{2}) = F(\frac{3}{2}) = \lambda/2$.

To solve the system (2) with Hamiltonian function (D.1), we use the same variables x and y as in the proof of Theorem 2. These variables are defined by

$$\begin{aligned} F(q(x)) &= \tfrac{1}{2}x^2 \text{ and } xq_f(x) \geq 0 \quad \text{for all} \quad x \in \mathbb{R}, \\ G(p(y)) &= \tfrac{1}{2}y^2 \text{ and } yq_g(y) \geq 0 \quad \text{for all} \quad y \in \mathbb{R}, \end{aligned}$$

as required by (42). We can check that $y = p$ and $x = \sqrt{1+2q} - 1$. Solving for q yields $q = q_f(x) = x + \frac{1}{2}x^2$, and hence (43) gives $t_f(x) = q_f{}'(x) = 1 + x$ and $t_g(y) = q_g{}'(y) = 1$.

System (50) for the phase angle in the (x, y)-plane takes the form

$$\frac{d\varphi}{dt} = -\frac{\sqrt{\lambda}}{1+\sqrt{2E}\cos\varphi}. \tag{D.2}$$

Integrating (D.3) over the range $\varphi \in [-\pi/2, \pi/2]$ gives the positive half period

$$T^{+} = T^{++} + T^{+-} = \frac{1}{\sqrt{\lambda}} \int_{-\pi/2}^{\pi/2} \left(1 + \sqrt{\frac{2E}{\lambda}} \cos\varphi\right) d\varphi = \frac{1}{\sqrt{\lambda}} \left(\pi + 2\sqrt{\frac{2E}{\lambda}}\right). \tag{D.3}$$

A similar integral over the range $\varphi \in [\pi/2, 3\pi/2]$ gives the negative half period

$$T^{-} = \frac{1}{\sqrt{\lambda}} \left(\pi - 2\sqrt{\frac{2E}{\lambda}}\right). \tag{D.4}$$

We see that both half periods depend on the energy E, but the total period $T = T^{+}(E) + T^{-}(E)$ is independent of E.

Now consider the energy-period function in case of the perturbed Hamiltonian

$$H(p, q) = H_0(p, q) - \varepsilon \frac{q^{n+1}}{n+1}$$

for a small perturbation parameter $\varepsilon > 0$. In the case $n \geq 3$, we get $f_1(0) = 0$, and hence (25b) implies $T'(0) = 0$. A numerical calculation for the example $\varepsilon = 0.1$ and $n = 3$ shows that $F_4(q) \geq 0$ holds for all $q \in [-0.3, \infty)$, but $f_1(q) \geq 0$ holds only for all $q \in [-0.2, \infty)$ and no longer for all $q \in [-0.3, -0.2]$. Hence we see—as follows from Theorem 1 as well—that Chow's functional F_4 implies increasing periods in a bit larger range of oscillation amplitudes than Schaaf's functional f_1.

6. The inverse problem. The following theorem considers the inverse problem to calculate the potential $F = F(q) \in \mathfrak{F}$ from a given energy-period function. To get a managable problem for which uniqueness is reasonable, we restrict ourselves to the system (6) with only one unknown potential $F = F(q)$. The inverse problem has been treated by Keller [7], too. It leads to an Abel integral equation, which can easily be solved by means of Laplace transformation already set up in this note. Hence we find it convenient to include this part. Beyond that, we get additional insights, as nessecary conditions for monotonicity of the energy-period function.

For given $E > 0$, let $q_- < 0 < q_+$ be the two solutions of $F(q_\pm) = E$, and define the function $v(E) = q_+ - q_-$. From (42), (43), and (79) we conclude

$$\frac{dv}{dE} = \frac{dv}{dx}\frac{dx}{dE} = 2s_f(x)\frac{dx}{dE} = 2\frac{s_f(\sqrt{2E})}{\sqrt{2E}} = \tau_f(E). \tag{110}$$

From now on, we omit the index f because we consider only one force function. Define the Abel integral operator $\mathfrak{A} : \tau \mapsto T$ by

$$T(E) = \int_0^\infty \sqrt{\tfrac{2}{a}}\, \tau(E-a)\, da = \left\{\sqrt{\tfrac{2}{E}}\right\} * \{\tau(E)\}. \tag{111}$$

In this operational notation, formula (95) for the energy-period function is simply $T = \mathfrak{A}\tau$ or $\{T(E)\} = \mathfrak{A}\{\tau(a)\}$.

THEOREM 4. *For the system (6), the energy-period function $T = T(E)$, the area $A = A(E)$ enclosed by orbits, and the Lagrangian $L = A - ET$ satisfy*

$$\begin{aligned} \{A(E)\} &= \mathfrak{A}\{2v(a)\}, & \{T(E)\} &= \mathfrak{A}\{\tau(a)\}, \\ \{ET(E) - \tfrac{1}{2}A(E)\} &= \mathfrak{A}\{a\tau(a)\}, & \{L(E)\} &= \mathfrak{A}\{v(a) - a\tau(a)\}, \\ \{T'(E)\} &= \mathfrak{A}\left\{\tfrac{d}{da}\left[\tau(a) - \sqrt{\tfrac{2}{f'(0)a}}\right]\right\} & \{ET'(E)\} &= \mathfrak{A}\{s'(\sqrt{2a})\}. \end{aligned}$$

Conversely, we can calculate the functions $v = v(a)$ and $v' = \tau$ from a given energy-period function by means of

$$\begin{aligned} \{v(a)\} &= \tfrac{1}{2\pi}\mathfrak{A}\{T(E)\}, \\ \left\{\tau(a) - \tfrac{1}{2\pi}T(0)\sqrt{\tfrac{2}{a}}\right\} &= \tfrac{1}{2\pi}\mathfrak{A}\{T'(E)\}, \\ \{a\tau(a) - \tfrac{1}{2}v(a)\} &= \tfrac{1}{2\pi}\mathfrak{A}\{ET'(E)\}, \\ \{a\tau'(a) + \tfrac{1}{2}\tau(a)\} &= \tfrac{1}{2\pi}\mathfrak{A}\{ET''(E) + T'(E)\}. \end{aligned} \tag{112}$$

By taking the Laplace transforms, these claims are rather straightforward to check. We leave the details to the reader.

Remarks.

(1) Obi [10] gives a criterion for constant periods of system (6), which we can derive from (113), too. Introducing Obi's function

$$E_{\mathrm{obi}}(x) = E_{\mathrm{obi}}(-x) = \frac{(q_+(x^2/2) + q_-(x^2/2))}{2} \quad \text{for } x \geq 0,$$

we get $q_+ + q_- = 2E_{\mathrm{obi}}\left(\sqrt{2F(q_+)}\right) = 2E_{\mathrm{obi}}\left(\sqrt{2F(q_-)}\right)$.

Constant periods $T(E) = 2\pi$ for all E occur if and only if $v(E) = 2\sqrt{2E}$, which is equivalent to $q_+ - q_- = 2\sqrt{2F(q_+)}$. Using Obi's function to eliminate either q_+ or q_-, we see that constant periods 2π occur if and only if Obi's condition

$$2F(q) = \left(q - E_{\mathrm{obi}}\left(\sqrt{2F(q)}\right)\right)^2$$

holds for all $q \in \mathbb{R}$.

(2) Under the additional symmetry assumption $F(q) = F(-q)$, formula (113) enables us to determine the potential from the given energy-period function, since the function $q^+ = \frac{1}{2}(q^+ - q^-) = \frac{1}{2}v(E)$ is the inverse of the potential $E = F(q^+)$.

(3) We get the following chain of implications among properties of the system (6). Roughly speaking, each step corresponds to a half-order integration achieved by the operator $\mathfrak{A}$.

$$\begin{aligned} &\frac{d}{dE}\left(\frac{E}{T}\frac{dT}{dE}\right) > 0 \quad \text{for all} \quad E > 0 \Longrightarrow \frac{ds}{da} > 0 \quad \text{for all} \quad a > 0 \Longrightarrow \\ &\qquad \frac{dT}{dE} > 0 \quad \text{for all} \quad E > 0 \Longrightarrow a\tau(a) - v(a) > 0 \quad \text{for all} \quad a > 0 \Longrightarrow \\ &\qquad L(E) > 0 \quad \text{for all} \quad E > 0. \end{aligned}$$

7. Acknowledgment. The author wants to thank the anonymous referees for useful suggestions, helping to clarify the connection to the bifurcation for two point boundary value problems, and a more complete list of references.

REFERENCES

[1] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, Berlin, 1978.

[2] R. Becker, *Theorie der Wärme*, Heidelberger Taschenbücher Bd. 10, Springer-Verlag, Berlin, Heidelberg, New York, 1975.

[3] C. Chicone, *The monotonicity of the period function for planar Hamiltonian vector fields*, J. Differential Equations, 69 (1987), pp. 310–321.

[4] S. N. Chow and D. Wang, *On the monotonicity of the period function of some second order equations*, Časopis Pést. Mat., 111 (1986), pp. 14–25.

[5] N. Chafee and E. Infante, *A bifurcation problem for a nonlinear parabolic equation*, J. Math. Anal. Appl., 4 (1974), pp. 17–37.

[6] J. K. Hale, *Dynamics in parabolic equations—an example*, Proceedings of the NATO Advanced Study Institute on Systems of Nonlinear Partial Differential Equations, J. M. Ball, ed., NATO ASI Series C, Math. and Phys. Sci., 111 (1982), pp. 461–472.

[7] J. B. Keller, *Inverse problems*, Amer. Math. Monthly, 83 (1976), pp. 107–118.

[8] W. S. Loud, *Behavior of the period of solutions. The behavior of the period of solutions of certain plane autonomous systems near centers*, Contr. Differential Equations, 3 (1964), pp. 21–36.

[9] J. Maynard-Smith and J. Hofbauer, *The "battle of the sexes," a genetic model with limit cycle behavior*, Theor. Pop. Biol., 32 (1987), pp. 1–14.

[10] Chike Obi, *Analytical theory of non-linear oscillations,* VIII *second order conservative systems whose solutions are all oscillating with the least period* 2π, Ann. Mat. Pur. Appl., 117 (1978), pp. 339–347.

[11] Z. Opial, *Sur les pèriodes des solutions de l'équation differentielle* $x'' + g(x) = 0$, Ann. Polon. Math., 10 (1961), pp. 49–72.

[12] F. Rothe, *The periods of the Volterra–Lotka system*, J. Reine Angew. Math., 355 (1985), pp. 129–138.

[13] ———, *Thermodynamics, real and complex periods of the Volterra model*, Z. Angew. Math. Phys., 36 (1985), pp. 395–421.

[14] R. Schaaf, *A class of Hamiltonian systems with increasing periods*, J. Reine Angew. Math., 363 (1985), pp. 96–109.

[15] ———, *Global Solution Branches of Two Point Boundary Value Problems*, Lecture Notes in Math., 1458, Springer-Verlag, Berlin, Heidelberg, New York, 1990.

[16] D. Wang, *The critical points of the period function of* $x'' - x^2(x - \alpha)(x - 1)(0 \leq \alpha < 1)$, Nonlinear Anal., 11 (1987), pp. 1029–1050.

[17] J. Waldvogel, *The period of the Volterra system is monotonic*, J. Math. Anal. Appl., 114 (1986), pp. 178–184.

SIAM J. MATH. ANAL.
Vol. 24, No. 1, pp. 155–171, January 1993

010

SOLUTIONS OF THIRD-ORDER DIFFERENTIAL EQUATIONS RELEVANT TO DRAINING AND COATING FLOWS*

WILLIAM C. TROY[†]

Abstract. The behavior of solutions of two differential equations is investigated. The first is a model for a viscous fluid draining over a wet surface. The second equation is derived from the first as a result of an inner expansion as a parameter δ tends to zero. The existence of the appropriate solution is proved for both equations.

Key words. differential equations, nonlinear, fluid mechanics, draining

AMS(MOS) subject classifications. 34B15, 34E05, 76O99

1. Introduction.

We investigate the behavior of solutions of two third-order differential equations, namely,

$$y''' = -1 + (1+\delta+\delta^2)\,y^{-2} - (\delta+\delta^2)\,y^{-3}, \qquad 0<\delta<1 \tag{1.1}$$

and

$$y''' = y^{-2} - y^{-3}. \tag{1.2}$$

Equation (1.1) arises in modeling draining or coating fluid flow problems. The function y represents the thickness of a thin film of viscous fluid draining over a wet solid surface in an unsteady manner. Equation (1.2) is the so-called "small δ limit" of (1.1), or inner expansion of (1.1) valid for $x \approx x_0$, obtained by setting $x = x_0 + \delta X, y = \delta Y$, followed by the formal limit $\delta \longrightarrow 0$.

In a recent article Tuck and Schwartz [3] present an extensive numerical study of the behavior of solutions of (1.1) and (1.2). The reader is referred to their paper for a more complete derivation of (1.1) and (1.2).

The physically important solution of (1.1) satisfies $(y, y', y'') \longrightarrow (\delta, 0, 0)$ as $x \longrightarrow \infty$, and $(y, y', y'') \longrightarrow (1, 0, 0)$ as $x \longrightarrow -\infty$. The numerical results of Tuck and Schwartz predict the existence of exactly one such solution, up to translation. We give a rigorous proof for the existence of at least one such solution in the first of our two main results (see Fig. 1).

THEOREM 1. *Let* $0 < \delta < 1$. *There is at least one solution* (*up to translation*) *of* (1.1) *satisfying* $\lim_{x\to\infty}(y, y', y'') = (\delta, 0, 0)$ *and* $\lim_{x\to-\infty}(y, y', y'') = (1, 0, 0)$.

Next we investigate the behavior of solutions of (1.2). Tuck and Schwartz show that the appropriate conditions to be satisfied by a solution of (1.2) are

$$y > 0 \quad \text{for all } x \in (-\infty, \infty)\,, \tag{1.3}$$

$$\lim_{x\to\infty}(y, y', y'') = (1, 0, 0)\,, \tag{1.4}$$

*Received by the editors November 5, 1990; accepted for publication (in revised form) February 25, 1992. This research was partially supported by National Science Foundation grant DMS-9002028.

[†] University of Pittsburgh, Department of Mathematics and Statistics, Pittsburgh, Pennsylvania 15260.

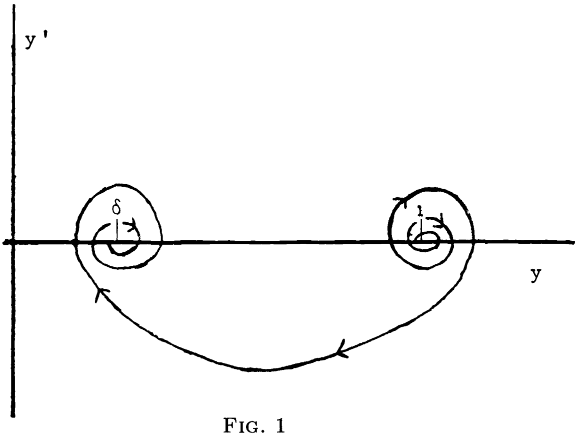


FIG. 1

and

$$\lim_{x \to -\infty} (y, y', y'') = (\infty, -\infty, 0). \tag{1.5}$$

Tuck and Schwartz also claim that any solution of (1.2)–(1.5) satisfies the asymptotic behavior

$$y \sim 3^{1/3}|x| \log^{1/3}(|x|) \quad \text{as } x \longrightarrow -\infty. \tag{1.6}$$

We confirm these properties in our second result (see Fig. 2).

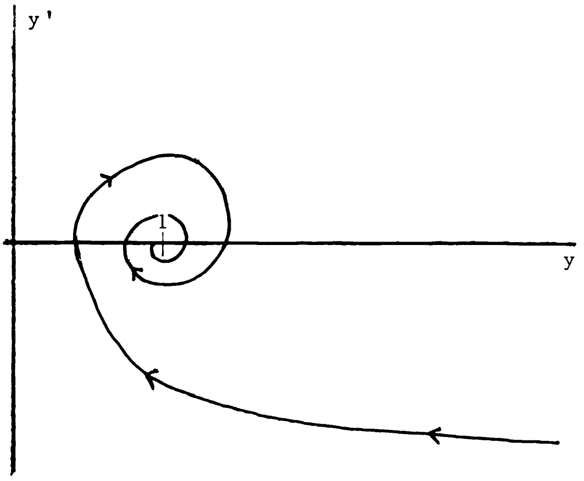


FIG. 2

THEOREM 2. *There exists a solution of* (1.2) *satisfying* (1.3)–(1.6).

2. Proof of Theorem 1. The proof uses a two-dimensional shooting argument. That is, we analyze the behavior of solutions of the initial value problem

$$y''' = -1 + (1+\delta+\delta^2)\,y^{-2} - (\delta+\delta^2)\,y^{-3}, \tag{2.1}$$

$$y(0) = \alpha, \quad y'(0) = \beta, \quad y''(0) = 0, \tag{2.2}$$

where $(\alpha,\beta) \in E \equiv \{(\alpha,\beta)\,|\delta \le \alpha \le 1 \text{ and } \beta \le 0\}$. For each $(\alpha,\beta) \in E$, let $[0,\bar{x})\,, \bar{x} = \bar{x}\,(\alpha,\beta)$ denote the maximal interval of existence of the solution of (2.1), (2.2) in $[0,\infty)$. Likewise, let $(\tilde{x},0]\,, \tilde{x} = \tilde{x}\,(\alpha,\beta)$ denote the maximal interval of existence of the solution in $(-\infty,0]$. It is possible that $\tilde{x}$ or $\bar{x}$ may be finite or infinite.

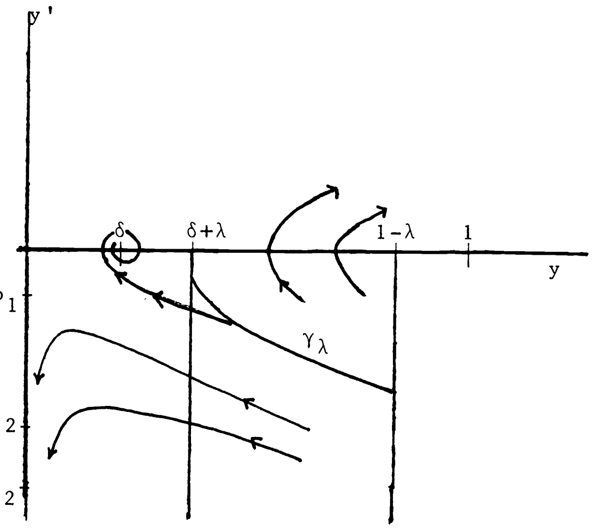


FIG. 3

In the first part of our proof we determine the behavior of solutions over $[0,\bar{x})$ for appropriately chosen values of α and β (see Fig. 3). In particular, for $\delta < \alpha \le 1$ and small $\beta < 0$, we find (Lemmas 1 and 3) that y' has a first zero on $[0,\bar{x})$, and $y(x_0) = \alpha$ for some first $x_0 > 0$. Next, in Lemmas 2 and 4 we show that if $\delta \le \alpha \le 1$ and $\beta < 0$ is large, then y' cannot have a zero and the solution enters an invariant set K_1 in which y' and y'' must remain negative. We then appeal to a topological result of McLeod and Serrin (Theorem 3) and show in Lemma 5 that there is a continuum γ_λ contained in $\{\delta + \lambda \le \alpha \le 1 - \lambda,\ \beta < 0\}$ for small $\lambda > 0$, which joins the lines $\alpha = \delta + \lambda$ and $\alpha = 1 - \lambda$, and such that if $(\alpha,\beta) \in \gamma_\lambda$, then $\lim_{x\to\infty}(y,y',y'') = (\delta,0,0)$. Next, we perform a similar analysis that determines the behavior of solutions on $(\tilde{x},0]$ (Lemmas 6–11).

We construct a second continuum, Γ_λ, joining the same lines as γ_λ such that if $(\alpha,\beta) \in \Gamma_\lambda$, then $\lim_{x\to-\infty}(y,y',y'') = (1,0,0)$ (see Fig. 4). For small $\lambda > 0$ we show that $\gamma_\lambda \cap \Gamma_\lambda \ne \phi$. Thus, if $(\bar{\alpha},\bar{\beta}) \in \gamma_\lambda \cap \Gamma_\lambda$ for small $\lambda > 0$, then Theorem 1 follows.

One of the main tools in our analysis is the energy functional

$$Q = y'y'' + y + \frac{(1+\delta+\delta^2)}{y} - \frac{(\delta+\delta^2)}{(2y^2)}, \tag{2.3}$$

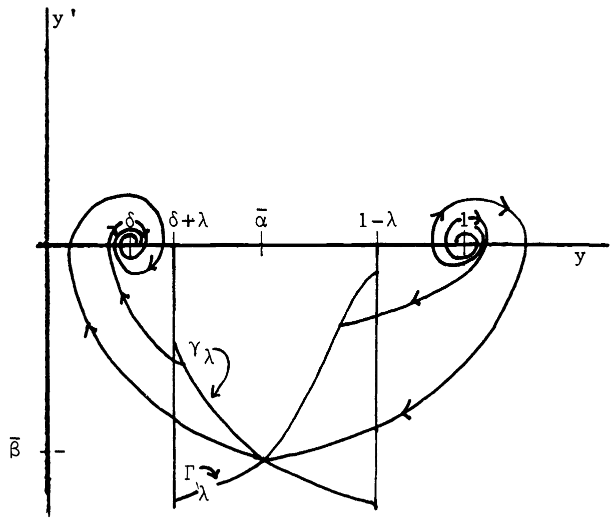


FIG. 4

which satisfies

$$Q' = (y'')^2. \tag{2.4}$$

We use Q to obtain our first technical result. Lemma 1 will play a key role in the construction of one of our shooting sets.

LEMMA 1. *Let $\delta < \alpha \le 1$ and $\beta < 0$. If there is a first $x_0 \in (0, \bar{x})$ for which $y(x_0) = \alpha$, then $y'(x_0) > 0$.*

Proof. Since $\beta < 0$, the definition of x_0 implies that $y'(x_0) \ge 0$. We assume, for the sake of contradiction, that $y'(x_0) = 0$. This and (2.3) lead to

$$Q(x_0) = Q(0). \tag{2.5}$$

For $\delta < \alpha < 1$, (2.1) gives $y'''(0) > 0$. Therefore, $y'' > 0$ on a small interval $(0, \epsilon)$, and an integration of (2.4) shows that

$$Q(x_0) = Q(0) + \int_0^{x_0} (y'')^2 dt \ge Q(0) + \int_0^{\epsilon} (y'')^2 dt > Q(0),$$

contradicting (2.5). If $\alpha = 1$, then $y'''(0) = 0$ and $y''''(0) = \beta(\delta^2 + \delta - 2) > 0$. Again $y''' > 0$ and $y'' > 0$ on a small interval $(0, \epsilon)$, and the same contradiction occurs. Thus we must conclude that $y'(x_0) > 0$.

Before defining our two shooting sets we need to determine the properties of a second type of solution, one which intersects the subset of R^3 given by

$$K_1 = \{(y, y', y'') \in R^3 | 0 < y < \delta, y' < 0, y'' < 0\}. \tag{2.6}$$

LEMMA 2. *If a solution of* (2.1), (2.2) *satisfies $(y(\hat{x}), y'(\hat{x}), y''(\hat{x})) \in K_1$ for some $\hat{x} \in [0, \bar{x})$, then $(y(x), y'(x), y''(x)) \in K_1$ for all $x \in [\hat{x}, \bar{x})$.*

Proof. It follows from (2.1) that $y'''(x) < 0, y'' < 0, y' < 0$, and $y > 0$, for all $x \in [\hat{x}, \bar{x})$.

We are now prepared to define our topological shooting sets by the following:

$A = \{(\alpha,\beta) \in E | y(x_0) = \alpha$ for some first $x_0 > 0$ and $0 < y(x) < \alpha$ on $(0,x_0)\}$;

$B = \{(\alpha,\beta) \in E | (y(\hat{x}), y'(\hat{x}), y''(\hat{x})) \in K_1$ for some $\hat{x} > 0$, and $0 < y(x) < \alpha$ for $0 < x < \hat{x}]\}$.

Remarks. We observe that $(y,y',y'') \equiv (\delta,0,0)$ and $(y,y',y'') \equiv (1,0,0)$ are solutions of (1.1). Thus, if $(\alpha,\beta) \in A \cup B$, then $(\alpha,\beta) \neq (\delta,0), (\alpha,\beta) \neq (1,0)$. Furthermore, if $(\alpha,\beta) = (\delta,\beta)$ and $\beta < 0$, then $y'''(0) = 0, y^{(4)}(0) < 0$, and (y,y',y'') immediately enters K_1. Thus $(\delta,\beta) \in B$ if $\beta < 0$.

It follows from Lemmas 1 and 2 and continuity that A and B are relatively open, disjoint subsets of E. We need to show that $A \neq \phi$ and $B \neq \phi$. For this we have the next two lemmas (see Fig. 3).

LEMMA 3. *For each $\alpha \in (\delta, 1]$ there exists $\beta_\alpha < 0$ such that if $\beta_\alpha < \beta < 0$, then $y(x_0) = \alpha$ for some first $x_0 > 0$.*

Proof. If $\alpha \in (\delta, 1)$ and $-\beta > 0$ is sufficiently small, then $y''' > 0$ for $x > 0$ until $y = 1$. This and continuity imply the result. If $\alpha = 1$ and $-\beta > 0$ is sufficiently small, then a linearization shows that (y,y') spirals around $(1,0)$ in the (y,y') plane, and the lemma follows.

It follows from Lemma 3 that $(\alpha,\beta) \in A$ if $\delta < \alpha \leq 1, \beta_\alpha < \beta < 0$.

LEMMA 4. *There exists $\tilde{\beta} \leq 0$ such that if $\delta \leq \alpha \leq 1$ and $\beta \leq \tilde{\beta}$, then $(y(\hat{x}), y'(\hat{x}), y''(\hat{x})) \in K_1$ for some $\hat{x} \in (0, \bar{x})$, and $y' < 0$ for $0 \leq x \leq \hat{x}$.*

Proof. Equation (2.1) shows that $|y'''|$ is bounded while $\delta/2 \leq y \leq 1$. Therefore, for large $-\beta > 0$, and $\delta \leq \alpha \leq 1$, there must be a first $x_1 \in (0, \bar{x})$ for which $y(x_1) = \delta/2$, and $y' < \beta + 1 << -2, y'' < 1$ for all $x \in (0, x_1]$. If $y''(x_1) \leq 0$, then $y'''(x_1) < 0$, and Lemma 2 shows that $(y,y',y'') \in K_1$ for all $x \in (x_1, \bar{x})$. We assume, for the sake of contradiction, that $y'' > 0$ for $x > x_1$ as long as $y' \leq 0$ and $y > 0$. Since (2.1) is autonomous, we may take $x_1 = 0$ so that

$$y(0) = \frac{\delta}{2}, \quad \beta < y'(0) < \beta + 1, \quad 0 < y''(0) < 1. \tag{2.7}$$

For $x > 0$, as long as $y' \leq 0$ and $y > 0$ we see that $y < \delta/2$ so that $y''' \leq -5$,

$$y'' \leq 1 - 5x, \quad \text{and} \quad y' \leq (\beta + 1) + x - \tfrac{5}{2}x^2. \tag{2.8}$$

Suppose for large $-\beta > 0$, that $y > 0$ over $[0, \frac{1}{5}]$ as long as $y'' > 0$. Then (2.8) shows that there is a first $\hat{x} \in \left(0, \frac{1}{5}\right]$ for which $y''(\hat{x}) = 0$, and $y' < 0$ on $[0, \hat{x})$. But then (y,y',y'') enters K_1 at $\hat{x}$. Thus the only possibility that remains is that $y(b) = 0$ for some first $b \in \left(0, \frac{1}{5}\right)$, and $y'' > 0, y' < 0$ on $[0, b)$. Let $0 < \epsilon < \delta/2$ be arbitrarily chosen, and let $a_2 > a_1 > 0$ satisfy $y(a_1) = \epsilon$ and $y(a_2) = \epsilon/2$. Then $y' > \beta$ over $[a_1, a_2]$ so that $a_2 - a_1 \geq -\epsilon/2\beta$. Also, over $[a_1, a_2]$ (2.1) gives $y''' < -(\delta + \delta^2)/2\epsilon^3$ for small $\epsilon > 0$. Therefore, $y'' < 1 - ((\delta + \delta^2)/2\epsilon^3)(x - a_1)$ over $a_1 < x < a_2$. In particular, we see that $y''(a_2) < 1 + ((\delta + \delta^2)/4\epsilon^2\beta) < 0$ for small $\epsilon > 0$, a contradiction. This completes the proof.

It follows from Lemma 4 that $(\alpha,\beta) \in B$ if $\delta \leq \alpha \leq 1$ and $\beta \leq \tilde{\beta}$. In order to complete the first part of our proof of Theorem 1 we need a topological result proved by McLeod and Serrin [1].

THEOREM 3 (see [1]). *Let I be the closed unit square $\{0 \leq x \leq 1, 0 \leq y \leq 1\}$ in the (x,y) plane, and let S^- and S^+ be disjoint relatively open subsets of I, respectively,* *containing the lines $y = 0$ and $y = 1$. Then the complement D of S^+ and S^- in I contains a continuum joining the lines $x = 0$ and $x = 1$.*

The McLeod–Serrin result applies to any closed rectangle in the plane. We need to apply this result to a subset of our set E in the (α, β) plane. For each small $\lambda > 0$ we define the subset E_λ of E by

$$E_\lambda = \{(\alpha,\beta) \in E | \delta + \lambda \leq \alpha \leq 1 - \lambda \text{ and } \tilde{\beta} \leq \beta \leq 0\}.$$

We define $S_\lambda^+ = A \cap E_\lambda$, $S_\lambda^- = B \cap E_\lambda$. Since A and B are relatively open in E, then S_λ^+ and S_λ^- are relatively open subsets of E_λ. Furthermore, Lemma 3 and continuity imply that the set $\delta + \lambda \leq \alpha \leq 1 - \lambda$, $\beta_\alpha < \beta < 0$ is contained in S_λ^+. Lemma 4 and continuity imply that the line segment $\beta = \tilde{\beta}$, $\delta + \lambda \leq \alpha \leq 1 - \lambda$ is contained in S_λ^-. Also, $S_\lambda^+ \cap S_\lambda^- = \phi$ since $A \cap B = \phi$. Thus, by Theorem 3 there is a continuum γ_λ contained in the complement $D_\lambda = E_\lambda - (S_\lambda^+ \cup S_\lambda^-)$ and which joins the lines $\alpha = \delta + \lambda, \tilde{\beta} < \beta < \beta_{\delta+\lambda}$ and $\alpha = 1 - \lambda$, $\tilde{\beta} < \beta < \beta_{1-\lambda}$. Recall that the entire line segment $\alpha = \delta$, $\tilde{\beta} < \beta < 0$ is contained in B. Similarly, by Lemma 3 the entire line segment $\alpha = 1$, $\beta_1 < \beta < 0$, is contained in A. The "endpoints" of γ_λ consist of the two sets

$$G_\lambda = \{(\delta + \lambda, \beta) |\ (\delta + \lambda, \beta) \in \gamma_\lambda\},$$

$$H_\lambda = \{(1 - \lambda, \beta) |\ (1 - \lambda, \beta) \in \gamma_\lambda\}.$$

Set $b_1(\lambda) \equiv \inf G_\lambda$ and $b_2(\lambda) \equiv \sup H_\lambda$. From these definitions, continuity, and our preceding discussion it follows that $b_1(\lambda) \longrightarrow 0$ as $\lambda \longrightarrow 0^+$, and $b_2(\lambda) \leq \beta_1/2 < 0$ for all small $\lambda > 0$. (See Fig. 3.) Next, we determine the behavior of solutions that satisfy $(\alpha, \beta) \in \gamma_\lambda$, where $\delta + \lambda < 1 - \lambda$.

LEMMA 5. *Let $(\alpha, \beta) \in \gamma_\lambda$. Then*

(i) $0 < \inf\{y(\zeta) | 0 < \zeta < \bar{x}\} \leq y(x) < \alpha$ *for* $0 < x < \bar{x}$;

(ii) $\bar{x} = \infty$;

(iii) $\lim_{x\to\infty} (y, y', y'') = (\delta, 0, 0)$.

Remarks. In the (y, y') plane a linearization shows that the solution spirals to $(\delta, 0)$ as $x \to \infty$. This and (i) imply that all relative minima of y remain bounded away from zero, while the relative maxima of y are less than α.

Proof. Suppose that $y(x_0) = \alpha$ for some first $x_0 > 0$. Then $(\alpha, \beta) \in A$, a contradiction. Next, we assume, for the sake of contradiction, that $\inf_{0<\zeta<\bar{x}}\{y(\zeta)\} = 0$. If, in fact, $\lim_{x\to\bar{x}} y(x) = 0$, then there is an $x_1 \in (0, \bar{x})$ with $y(x_1) = \delta/2$, and $y \in (0, \delta/2)$ for all $x \in (x_1, \bar{x})$. It then follows exactly as in the last part of the proof of Lemma 4 that (y, y', y'') eventually enters K_1. Thus $(\alpha, \beta) \in B$, again a contradiction. The remaining alternative is that $0 = \underline{\lim}_{x\to\bar{x}}\, y(x) < \overline{\lim}_{x\to\bar{x}}\, y(x)$. Then there exists an increasing sequence $\{x_N\}_N$ of positive numbers with $\lim_{N\longrightarrow\infty} x_N = \bar{x}, y'(x_N) = 0$ for all N, and $\lim_{N\to\infty} y(x_N) = 0$. This and (2.3) imply that $Q(x_N) \longrightarrow -\infty$ as $N \longrightarrow \infty$, a contradiction since $Q(0) > -\infty$ and $Q' \geq 0$ for all $x \in [0, \bar{x})$. Thus we must conclude that $\inf\{y(\zeta) | 0 < \zeta < \bar{x}\} > 0$, and (i) is proved. With y bounded below away from zero and bounded above by α, it follows from (2.1) that

$$|y'''| < M \quad \text{for } 0 < x < \bar{x} \tag{2.9}$$

for some $M > 0$. An integration of (2.9) shows that y' and y'' are bounded over any finite interval. Therefore, (y', y'') cannot become unbounded at any finite value, and

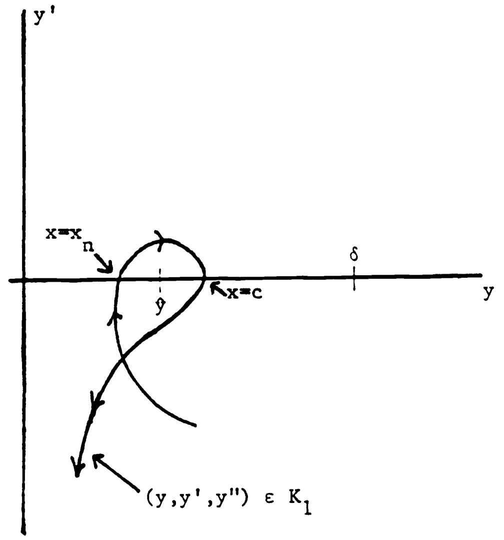


FIG. 5

it must be the case that $\bar{x} = \infty$ so that (ii) is proved. It remains to be shown that $\lim_{x\to\infty}(y, y', y'') = (\delta, 0, 0)$. First we show that Q is bounded above. If, in fact, $Q \longrightarrow \infty$ as $x \longrightarrow \infty$, it must be the case that $\lim_{x\to\infty} y'y'' = \infty$ since y is bounded away from zero. This leads either to $y \longrightarrow \infty$ or $y \longrightarrow -\infty$, which, by part (i), cannot be true. Therefore, $\lim_{x\to\infty} Q(x) = \bar{Q} < \infty$. Suppose that $\underline{\lim}_{x\to\infty} y'' = -\mu < 0$. Since y is bounded, it must be the case that $\overline{\lim}_{x\to\infty} y'' \geq 0$. Therefore, there is an unbounded increasing sequence $\{x_N\}_N$ such that $y''(x_N) = -\mu/2$ and $y'''(x_N) \geq 0$ for all N. Choose N such that

$$Q(x_N) > \bar{Q} - \frac{\mu^3}{64M}. \tag{2.10}$$

Let $a = x_N$, and let $b > a$ the succeeding point for which $y(b) = -\mu/4$. Then $y''' \leq M$ on (a, b) so that $b - a \geq \mu/(4M)$. This and (2.4), (2.10) lead to $Q(b) \geq Q(a) + \mu^3/64M > \bar{Q}$, a contradiction. It follows that $\underline{\lim}_{x\to\infty} y'' \geq 0$. Similarly, $\lim_{x\to\infty} y'' \leq 0$ so that $\lim_{x\to\infty} y'' = 0$. Next, suppose that $\overline{\lim}_{x\to\infty} y(x) \neq \delta$. There are two possibilities. First, suppose that $\lim_{x\to\infty} y(x) = \bar{y}$ exists with $\bar{y} \in (0, \delta) \cup (\delta, \alpha]$. Then $\lim_{x\to\infty} y''' = \gamma \neq 0$, and an integration leads to $y''(\infty) \neq 0$, a contradiction. The other possibility is that $\underline{\lim}_{x\to\infty} y(x) < \overline{\lim}_{x\to\infty} y(x)$. First, suppose that $\underline{\lim}_{x\to\infty} y(x) = \hat{y} < \delta$. Let $\{x_N\}_N$ be an increasing unbounded sequence with $y'(x_N) = 0$ for all N, and $\lim_{N\to\infty} y(x_N) = \hat{y}$. Also, $\lim_{N\to\infty} y''(x_N) = 0$. From (2.1) there is an $\eta > 0$ such that $y''' \leq -\eta$ when $y \leq (\hat{y} + \delta)/2$. Thus, if $y''(x_n) > 0$, an integration of (2.1) from x_N to x shows that for large $N, y' = 0$ and $y'' \leq 0$ at some first $x = c > x_N$ before $y = \hat{y}/2$ (Fig. 5). Thus (y, y', y'') enters K_1 at $x = c$ so that $(\alpha, \beta) \in B$, a contradiction. Similarly, (y, y', y'') enters K_1 at x_N if $y''(x_N) < 0$. Therefore, $\underline{\lim}_{x\to\infty} y(x) \geq \delta$. If we assume that $\overline{\lim}_{x\to\infty} y(x) > \delta$, then similar reasoning shows that y exceeds α, again a contradiction. Therefore, $\lim_{x\to\infty} y(x) = \delta$. Finally, it remains to prove that $\lim_{x\to\infty} y' = 0$. If $\underline{\lim}_{x\to\infty} y' < 0$, then it easily follows from the fact that $y'' \to 0$ and an integration that $\underline{\lim}_{x\to\infty} y(x) < \delta$, a contradiction. Thus $\underline{\lim}_{x\to\infty} y' \geq 0$. Likewise, $\overline{\lim}_{x\to\infty} y' \leq 0$. Therefore, $\lim_{x\to\infty}(y, y', y'') = (\delta, 0, 0)$, and (iii) is proved.

We now proceed with the second part of our shooting argument in which we follow solutions backwards from zero. Again, let $\lambda > 0$ satisfy $\delta + \lambda < 1 - \lambda$. For each such λ, our goal is to prove that there is a second continuum $\Gamma_\lambda \subseteq E_\lambda$ that joins the lines $\alpha = \delta + \lambda, \beta < 0$, and $\alpha = 1 - \lambda, \beta < 0$ and has the following additional properties. (Refer to Fig. 4.)

LEMMA 6. *Let* $(\alpha, \beta) \in \Gamma_\lambda$;

(i) $\alpha < y(x)$ *for* $\tilde{x} < x < 0$;

(ii) $\tilde{x} = -\infty$;

(iii) $\lim_{x \to -\infty}(y, y', y'') = (1, 0, 0)$.

Furthermore, if $b_3 \equiv \sup\{\beta < 0 | (\delta + \lambda, \beta) \in \Gamma_\lambda\}$ *and* $b_4 \equiv \inf\{\beta < 0 | (1 - \lambda, \beta) \in \Gamma_\lambda\}$, *then* b_3 *is bounded away from zero for small* λ, *and* $b_4 \to 0$ *as* $\lambda \to 0$.

We postpone the proof of Lemma 6 in order to see its consequences (Fig. 4). It follows from the properties of γ_λ and Γ_λ given above that $\gamma_\lambda \cap \Gamma_\lambda \neq \phi$ for small $\lambda > 0$. Let $(\bar{\alpha}, \bar{\beta}) \in \gamma_\lambda \cap \Gamma_\lambda$. Lemmas 5 and 6 imply that the solution with $y(0) = \bar{\alpha}, y'(0) = \bar{\beta}, y''(0) = 0$ must exist for all $x \in (-\infty, \infty)$, that $\lim_{x \to -\infty}(y, y', y'') = (1, 0, 0)$ and $\lim_{x \to \infty}(y, y', y'') = (\delta, 0, 0)$. Theorem 1 follows. Thus it remains to prove the existence of Γ_λ. For this we find it convenient to set $t = -x$ so that the problem (2.1), (2.2) becomes

$$y''' = 1 - \frac{(1 + \delta + \delta^2)}{y^2} + \frac{(\delta + \delta^2)}{y^3} \tag{2.11}$$

with initial condition

$$y(0) = \alpha, \quad y'(0) = -\beta > 0, \quad y''(0) = 0, \tag{2.12}$$

where $' \equiv d/dt$. Also, we set $\tilde{t} = -\tilde{x}$. It is evident that Lemma 6 is equivalent to the following.

LEMMA 7. *Let* $(\alpha, \beta) \in \Gamma_\lambda$. *Then*

(i) $\alpha < y(t)$ *for* $0 < t < \tilde{t}$;

(ii) $\tilde{t} = \infty$;

(iii) $\lim_{t \to \infty}(y, y', y'') = (1, 0, 0)$.

The existence of Γ_λ follows from a two-dimensional shooting argument that is similar to that used to prove the existence of γ_λ. Again, the first step is to define an energy functional analogous to Q, namely,

$$P \equiv y'y'' - y - \frac{(1 + \delta + \delta^2)}{y} + \frac{(\delta + \delta^2)}{2y^2}, \tag{2.13}$$

which satisfies

$$P' = (y'')^2. \tag{2.14}$$

The existence of Γ_λ and the proof of Lemma 7 requires the next four lemmas, which are analogous to Lemmas 1–4.

LEMMA 8. *Let* $\delta < \alpha \leq 1$ *and* $-\beta > 0$. *If there is a first* $t_0 > 0$ *for which* $y(t_0) = \alpha$, *then* $y'(t_0) < 0$.

Proof. $y'(t_0) \leq 0$ by definition of t_0. Suppose that $y'(t_0) = 0$. Then (2.13) implies that

$$P(t_0) = P(0). \tag{2.15}$$

However, $-\beta > 0$ and $\delta < \alpha \leq 1$ so that (2.11), (2.12) imply that $y'' \neq 0$ on an interval $(0, \epsilon)$. Therefore, an integration of (2.14) gives $P(t_0) > P(0)$, contradicting (2.15).

For our next result, we define the set

$$(2.16) \qquad H_1 = \{(y, y', y'') \in R^3 | y > 1, y' > 0, y'' > 0\}.$$

The set H_1 will play a role similar to that of the set K_1 defined in (2.6). In Lemma 2 we described the crucial properties of K_2, and we now describe the key properties of H_1 in our next result.

LEMMA 9. *If* $\left(y\left(\hat{t}\right), y'\left(\hat{t}\right), y''\left(\hat{t}\right)\right) \in H_1$ *for some* $\hat{t}$, *then* $(y(t), y'(t), y''(t)) \in H_1$ *for all* $t \in \left(\hat{t}, \tilde{t}\right)$.

Proof. It follows from (2.11) that $y''' > 0, y'' > 0, y' > 0$, and $y > 1$ for all $t \in \left(\hat{t}, \tilde{t}\right)$.

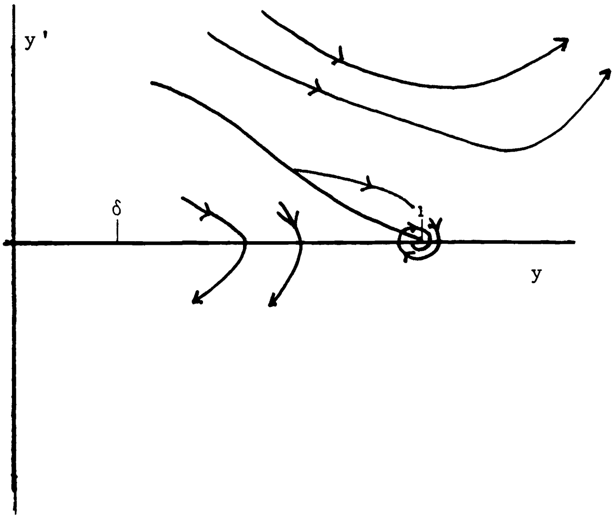


FIG. 6

We now define our topological shooting sets by (see Fig. 6)

$$C = \{(\alpha, \beta) \in E | \text{ there exists a first } t_0 > 0 \text{ with } y(t_0) = \alpha \text{ and } y > \alpha \text{ on } (0, t_0)\},$$
$$D = \{(\alpha, \beta) \in E | \left(y\left(\hat{t}\right), y''\left(\hat{t}\right), y''\left(\hat{t}\right)\right) \in H_1 \text{ for some } \hat{t} > 0,$$
$$\text{and } y > \alpha \text{ for all } t \in \left(0, \hat{t}\right]\}.$$

Remarks. It follows from the definitions of the sets C and D so that $(\alpha, \beta) \notin C \cup B$ if $(\alpha, \beta) = (\delta, 0)$ or $(\alpha, \beta) = (1, 0)$. Furthermore, if $(\alpha, \beta) = (1, \beta)$ and $\beta < 0$, then it follows from (2.11) that $y'''(0) = y''(0) = 0, y'(0) = -\beta > 0$ and $y^{(4)}(0) > 0$. Thus (y, y', y'') enters H_1 as t increases from zero, and we conclude that $(1, \beta) \in D$ for all $\beta < 0$.

It follows from Lemmas 8, 9, and continuity that C and D are both open in E, and $C \cap D = \phi$. In order to see that C and D are nonempty, we need the next two results. (See Fig. 6.) These are similar to Lemmas 3 and 4.

LEMMA 10. *For each $\alpha \in [\delta, 1)$ there exists $\beta^{\alpha} < 0$ such that if $-\beta^{\alpha} > -\beta > 0$, then $y(t_0) = \alpha$ for some first $t_0 > 0$.*

Proof. If $\alpha \in (\delta, 1)$ and $-\beta > 0$ is sufficiently small, then (2.11) implies that $y''' < 0$ for $t > 0$ until $y < \delta$. This and continuity imply the result. If $\alpha = \delta$ and $-\beta > 0$ is small, then a linearization of (2.11) around the constant solution $(y, y', y'') \equiv (\delta, 0, 0)$ shows that (y, y') spirals around $(\delta, 0)$ in the (y, y') plane, and the existence of t_0 is assured.

LEMMA 11. *There exists $\bar{\beta} < 0$ such that if $-\beta \geq -\bar{\beta}$ and $\delta \leq \alpha \leq 1$, then*

$$(y(\hat{t}), y'(\hat{t}), y''(\hat{t})) \in H_1 \quad \textit{for some } \hat{t} > 0, \quad \textit{and} \quad y'(t) > 0 \quad \textit{for all } t \in [0, \hat{t}].$$

Proof. Equation (2.11) shows that $|y'''|$ is bounded while $y \geq \alpha$. Thus, for large $-\beta > 0$ there is a first $t_1 \in (0, \bar{t})$ independent of $\alpha \in [\delta, 1]$ such that $y(t_1) = 2$, and $y' > -\beta - 1 > 2, y'' > -1$ for $0 \leq t \leq t_1$. If $y''(t_1) \geq 0$, then $y'''(t_1) > 0$; it follows from (2.11) that $(y, y', y'') \in H_1$ for $t > t_1$. We assume, for the sake of contradiction, that $y'' < 0$ for $t > t_1$ as long as $y' > 0$. Since (2.11) is autonomous, we may set $t_1 = 0$ and consider the initial values

$$y(0) = 2, \quad 2 < -\beta - 1 < y'(0) < -\beta, \quad -1 < y''(0) < 0. \tag{2.17}$$

For $t > 0$, as long as $y' > 0$, equation (2.11) implies that $y''' \geq \frac{1}{2}$,

$$y'' \geq -1 + \frac{t}{2}, \qquad y' \geq 2 - t + \frac{t^2}{4}. \tag{2.18}$$

It easily follows from (2.18) that, if $-\beta > 0$ is sufficiently large, $y'' = 0$ at some $\bar{t} \in (0, 2]$, and $y' > 0$ for all $t \in [0, \bar{t}]$. But then $y'''(\bar{t}) \geq \frac{1}{2}$; hence (y, y', y'') enters H_1 at $\bar{t}$.

We define $U_{\lambda}^{+} = C \cap E_{\lambda}$ and $U_{\lambda}^{-} = D \cap E_{\lambda}$, where, once again, $\lambda > 0$ satisfies $\delta + \lambda < 1 - \lambda$.

It follows from Lemma 10 that $(\alpha, \beta) \in U_{\lambda}^{+}$ if $\delta \leq \alpha < 1$ and $\beta^{\alpha} < \beta < 0$. Lemma 11 shows that $(\alpha, \beta) \in U_{\lambda}^{-}$ if $\delta \leq \alpha \leq 1$ and $|\beta - \bar{\beta}|$ is small. Also, $U_{\lambda}^{+} \cap U_{\lambda}^{-} = \phi$ since $C \cap D = \phi$. Thus, as before, we conclude from the result of McLeod and Serrin that there is a continuum $\Gamma_{\lambda} \subseteq E_{\lambda} - (U_{\lambda}^{+} \cup U_{\lambda}^{-})$ that joins the line segment $\alpha = \delta + \lambda, \bar{\beta} < \beta < 0$ with the line segment $\alpha = 1 - \lambda, \bar{\beta} < \beta < 0$. As with γ_{λ}, we investigate the endpoints of Γ_{λ}, which consist of the sets

$$J_{\lambda} = \{(\delta + \lambda, \beta) |\ (\delta + \lambda, \beta) \in \Gamma_{\lambda}\},$$
$$K_{\lambda} = \{(1 - \lambda, \beta) |\ (1 - \lambda, \beta) \in \Gamma_{\lambda}\}.$$

From continuity and the fact that the line segment $(1, \beta) \in D$ for $\beta < 0$ it follows that $\lim_{\lambda \to 0} \Gamma_{\lambda} \cap \{\alpha = 1 - \lambda\} = (1, 0)$. Furthermore, $(\delta, \beta) \in C$ for $\beta_{\delta} < \beta < 0$ so that $\sup J_{\lambda} < \beta_{\delta}/2$ for all small $\lambda > 0$. This, together with the properties of γ_{λ} described earlier, imply that $\gamma_{\lambda} \cap \Gamma_{\lambda} \neq \phi$ for sufficiently small $\lambda > 0$. Let $(\bar{\alpha}, \bar{\beta}) \in \gamma_{\lambda} \cap \Gamma_{\lambda}$ for small $\lambda > 0$. Then Lemmas 5 and 6 imply Theorem 1.

It remains to prove Lemma 7, which in turn implies Lemma 6. The proof of Lemma 7 uses the same arguments as the proof of Lemma 5. The main difference is that the functional P replaces the functional Q. For the sake of brevity, we omit a repetition of the details.

3. Proof of Theorem 2. Once again, we employ a two-dimensional shooting argument. Here the appropriate initial value problem is

$$y''' = y^{-2} - y^{-3}, \tag{3.1}$$

$$y(0) = \alpha, \quad y'(0) = \beta, \quad y''(0) = 0, \tag{3.2}$$

where $(\alpha, \beta) \in G \equiv \{(\alpha,\beta) | \frac{1}{2} \le \alpha \le 1, \beta \ge 0\}$. Let $[0, \bar{x})$ and $(\tilde{x}, 0]$ denote the maximal positive and negative intervals of existence of the solution of (3.1), (3.2). As in the previous section, we will make use of an energy functional, namely,

$$R = y'y'' + y^{-1} - \frac{y^{-2}}{2}, \tag{3.3}$$

which satisfies

$$R' = (y'')^2 \quad \text{for } \tilde{x} < x < \bar{x} \tag{3.4}$$

and

$$0 \le R(0) \le \tfrac{1}{2} \quad \text{for all } (\alpha, \beta) \in G. \tag{3.5}$$

As in §2 we need two technical lemmas before defining our shooting sets.

LEMMA 12. *For each $\alpha \in \left[\frac{1}{2}, 1\right)$ there exists $\beta_1(\alpha) > 0$ such that if $0 < \beta < \beta_1(\alpha)$, then $y(x_0) = \alpha$ for some first $x_0 > 0$, and $y'(x_0) < 0$.*

Proof. If $\beta = 0$, then $y'(0) = y''(0) = 0$ and $y'''(0) < 0$. This and continuity give the existence of x_0 for small $\beta > 0$. The definition of x_0 implies that $y'(x_0) \le 0$. It follows from (3.3) and (3.4) that $R(x_0) > R(0)$ so that $y'(x_0) < 0$.

Next, we define the subset K_2 of the phase space R^3 by

$$K_2 = \{(y, y', y'') \in R^3 | y > 1, y' > 0, y'' > 0\}. \tag{3.6}$$

LEMMA 13. *If $(y(\hat{x}), y'(\hat{x}), y''(\hat{x})) \in K_2$ for some $\hat{x} > 0$, then $(y(x), y'(x), y''(x)) \in K_2$ for all $x \in [\hat{x}, \bar{x})$.*

Proof. Equation (3.1) and the definition of K_2 imply that $y''' > 0, y'' > 0, y' > 0, y > 1$ for all $x \in [\hat{x}, \bar{x})$, and the lemma immediately follows.

We define our topological shooting (Fig. 6) sets by

$$\begin{aligned} X =& \{(\alpha, \beta) \in G | y(x_0) = \alpha \text{ for some first } x_0 > 0, \text{ and } y > \alpha \text{ for } 0 < x < x_0\}, \\ Y =& \{(\alpha, \beta) \in G | (y(\hat{x}), y'(\hat{x}), y''(\hat{x})) \in K_2 \text{ for some } \hat{x} > 0, \\ & \text{and } y(x) > \alpha \text{ for } 0 < x < \hat{x}\}. \end{aligned} \tag{3.7}$$

It follows from Lemma 12 and continuity that $X \ne \phi$ and X is open in G. Also, continuity implies that Y is open in G. Lemma 13 and the definitions of X and Y imply that $X \cap Y = \phi$. Finally, we show that $Y \ne \phi$ in the following lemma.

LEMMA 14. *There exists $\beta^{**} > 0$ such that if $\frac{1}{2} \leq \alpha \leq 1$ and $\beta \geq \beta^{**}$, then $(y(\hat{x}), y'(\hat{x}), y''(\hat{x})) \in K_2$ for some $\hat{x} \in (0, \bar{x})$, and $y' > 0$ for $0 \leq x < \bar{x}$.*

Proof. Equation (3.1) implies that $|y'''|$ is bounded as long as $y' > 0$. Thus, if $\beta >> 1$ and $\frac{1}{2} \leq \alpha \leq 1$, there is a first $x_1 = x_1(\alpha, \beta) \in (0,1)$ such that $y(x_1) = 2, y'(x) \geq \beta - 1 > 0$ and $y''(x) \geq -1$ over $[0, x_1]$. If $y''(x_1) \geq 0$, then $y'''(x_1) = \frac{1}{8}$, and it follows from Lemma 13 that $(y, y', y'') \in K_2$ for $x_1 < x < \bar{x}$. Suppose, however, that $y''(x) < 0$ for $0 < x < \bar{x}$ as long as $y' > 0$. In order to obtain a contradiction, we need a lower bound on $R(x_1)$. If $\frac{3}{4} \leq y(0) \leq 1$, then $R(0) \geq 12/27$. Thus, by (3.4), $R(x_1) \geq 12/27$. If $\frac{1}{2} \leq y(0) \leq \frac{3}{4}$, we let $0 < a_1 < a_2 < x_1$ satisfy $y(a_1) = \frac{3}{4}$ and $y(a_2) = \frac{7}{8}$. Then $y' < \beta$ and an integration over $[a_1, a_2]$ leads to $a_2 - a_1 \geq 1/(8\beta)$. Also, over $[0, a_2]$ it follows from (3.1) that $y''' \leq -64/343$. Thus $y'' \leq -64_x/343$, and we conclude from (3.43) that $R' \geq (64/343)^2 x^2$ and $R(x) \geq (64/343)^2 x^3/3$. At $x = x_1$, we obtain $R(x_1) \geq R(a_1) \geq 8/(3(343)^2\beta^3)$. Since our system is autonomous we may set $x_1 = 0$ and consider the initial value problem

$$(3.8) \qquad y(0) = 2, \quad \beta - 1 < y'(0) < \beta, \quad -1 \leq y''(0) < 0, \quad R(0) > \frac{8}{3(343)^2\beta^2}.$$

Suppose that $y' > 0$ for all $x \geq 0$. Then $y(\infty)$ exists with $2 < y(\infty) \leq \infty$. If $y(\infty) < \infty$, then $y''' \geq k \equiv (1/(y(\infty))^2) - (1/(y(\infty))^3) > 0$ on $[0, \infty)$. An integration shows that $y'' = 0$ at some first $x_1 > 0$. At $x_1, y(x_1) > 2, y'(x_1) > 0, y''(x_1) = 0$, and $y'''(x_1) > k > 0$ so that (y, y', y'') enters K_2 at x_1. Otherwise, if $y(\infty) = \infty$, then $R(\infty) > 0$, and we conclude from the definition of R that $y'y'' > 0$ for large x. Thus, once again, $y''(x_1) = 0$ at some first $x_1 > 0$ so that (y, y', y'') again enters K_2 at x_1. Finally, we must consider the possibility that $y'(a) = 0$ at some first $a > 0$. Since $y' > 0$ and $R > 0$ on(o, a),then $y'(y'y'' + (1/y) - (1/2y^2)) > 0$, and an integration gives $((y')^3/3 + \ln(y) + (1/2y) > \beta^3/4$ for large β. In particular, since $y(a) > 2$, then at $x = a$ we obtain $\ln(y(a)) > \beta^3/6$ for large β. That is, $y(a) > e^{\beta^3/6}$ for $\beta >> 1$. But then $R(a) \leq e^{-\beta^3/6} < 8/(3(343)^2\beta^3)$ for large β, contradicting (3.8). Thus a cannot exist for large β and the lemma is proved.

The next step in our analysis is to once again apply the McLeod–Serrin result. Let $\lambda > 0$ such that $\frac{1}{2} < 1 - \lambda$. Then, by Lemmas 12–14 and Theorem 3, there is a continuum g_λ contained in the set $\frac{1}{2} \leq \alpha \leq 1 - \lambda, \beta > 0$, which joins the line segments $\alpha = \frac{1}{2}, 0 < \beta < \beta^{**}$ and $\alpha = 1 - \lambda, 0 < \beta < \beta^{**}$ and such that $g_\lambda \cap (X \cup Y) = \phi$. Furthermore, since the line $\alpha = 1, \beta > 0$ is contained in the set Y, then we conclude that $\lim_{\lambda \to 0^+} g_\lambda \cap \{\alpha = 1 - \lambda\} = (1, 0)$. In order to complete the proof of Theorem 2 we need to prove two further results. First, we show in Lemma 15 that if $(\alpha, \beta) \in g_\lambda$, then $(y, y', y'') \longrightarrow (1, 0, 0,)$ as $x \longrightarrow \infty$. Following that, the rest of the proof is devoted to showing that there is a point $(\hat{\alpha}, \hat{\beta}) \in g_\lambda$ for which (1.5) and (1.6) hold.

LEMMA 15. *Let $(\alpha, \beta) \in g_\lambda$. Then*

(i) $\bar{x} = \infty$;

(ii) $\alpha < y(x) < \sup\{y(\zeta) | 0 \leq \zeta < \infty\} < \infty$;

(iii) $\lim_{x \to \infty}(y, y', y'') = (1, 0, 0)$.

Proof. If there were a first $x_0 > 0$ for which $y(x_0) = \alpha$, then Lemma 12 gives $y'(x_0) < 0$ so that $(\alpha, \beta) \in X$. This is a contradiction since $g_\lambda \cap (X \cup Y) = \phi$. Thus $y(x) > \alpha$ for all $x \in (0, \bar{x})$. Equation (3.1) shows that $|y'''|$ is bounded for all $x \in (0, \bar{x})$. Therefore, neither y' nor y'' can become unbounded at finite x. Also, since $y > \alpha$ for all $x \in (0, \bar{x})$, our conclusion must be that $\bar{x} = \infty$. This proves (i) and the first part of (ii). Next, suppose that $\sup\{y(\zeta) | 0 \leq \zeta < \infty\} = \infty$. Then $R(\infty) > 0$. If

y oscillates infinitely often as $x \longrightarrow \infty$, there is an unbounded, increasing sequence $\{x_i\}_i$ with $y(x_i) \longrightarrow \infty$ as $i \longrightarrow \infty$, and $y'(x_i) = 0$, all i. Then $\lim_{i\to\infty} R(x_i) = 0$, a contradiction. Thus it must be the case that $y' > 0$ and $y'' < 0$ for large x, and $y(x) \longrightarrow \infty$ as $x \longrightarrow \infty$. Again, $\lim_{x\to\infty} R(x) > 0$ so that $\lim_{x\to\infty} y'y'' > 0$. But then $y'' > 0, y' > 0, y > 1$ for large x and $(y, y', y'') \in K_2$, a contradiction. Therefore, $\sup\{y(\zeta)|0 \le \zeta < \infty\} < \infty$ and (ii) is proved. Part (iii) now easily follows from energy arguments similar to those used in Lemma 4. We omit the details for the sake of brevity.

Since (1.5) and (1.6) describe the behavior of the required solution as $x \longrightarrow -\infty$, we find it mathematically convenient to "turn the problem around" by setting $t = -x$. Then the appropriate initial value problem is

$$y''' = y^{-3} - y^{-2}, \tag{3.9}$$

$$y(0) = \alpha, \quad y'(0) = -\beta, \quad y''(0) = 0, \tag{3.10}$$

where $(\alpha, \beta) \in G$ and $' \equiv d/dt$. We need to determine the behavior of the solution of (3.9), (3.10) for $t \ge 0$, where $(\alpha, \beta) \in g_\lambda$. First we prove the following result for solutions satisfying $(\alpha, \beta) \approx (1, 0)$.

LEMMA 16. *There are values $\alpha_* \in (\frac{1}{2}, 1)$ and $\beta_* > 0$ such that if $\alpha_* < \alpha < 1$ and $0 \le \beta < \beta_*$, then $y''(t_0) = 0$ at some first $t_0 > 0$. Furthermore, $y''(t) > 0$ over $(0, t_0), y(t_0) > 1, y'(t_0) > 0$ and $y'''(t_0) < 0$.*

Proof. A linearization of (3.9) around the constant solution $y \equiv 1$ shows that the solution must spiral around $(1, 0)$ in the (y, y') plane if $1 - \alpha > 0$ and $\beta \ge 0$ are sufficiently small. In particular, since $\alpha < 1$, (3.19) implies that $y''' > 0$ for $t \ge 0$ until $y = 1$ at some first $\hat{t} > 0$. Here $y'(\hat{t}) > 0$ and $y''(\hat{t}) > 0$. Thus $y(t_0) > 1$, and it follows from (3.9) that $y'''(t_0) < 0$. This proves the lemma.

Remark. Recall that $\lim_{\lambda\to 0} g_\lambda \cap \{\alpha = 1 - \lambda\} = (1, 0)$. This and Lemma 16 immediately imply the following.

LEMMA 17. *If $\lambda > 0$ is small and $(1 - \lambda, \beta) \in g_\lambda$ for some $\beta > 0$, then there is a first $t_0 > 0$ such that $y''(t_0) = 0$. Furthermore, $y(t_0) > 1, y'(t_0) > 0, y'''(t_0) < 0$.*

We are now prepared to finish the proof of Theorem 2. Since g_λ is a continuum, we conclude that there is a continuous function $f : [\frac{1}{2}, 1 - \lambda] \longrightarrow (0, \infty)$ such that $(\alpha, f(\alpha)) \in g_\lambda$ for all $\alpha \in [\frac{1}{2}, 1 - \lambda]$. Furthermore, we assume that $\lambda > 0$ is sufficiently small so that Lemma 17 holds. We need to determine the behavior of the solution of $(3.9), (3.10)$, where $\beta = f(\alpha)$ as α increases from $\frac{1}{2}$ to $1 - \lambda$. That is, $y(0) = \alpha$ and $(dy/dt)(0) = -f(\alpha)$. (See Fig. 7.) We will use the functional

$$S \equiv y'y'' + \frac{y^{-2}}{2} - y^{-1}, \tag{3.11}$$

which satisfies

$$S' = (y'')^2 \tag{3.12}$$

and

$$-\tfrac{1}{2} \le S(0) \le 0 \quad \textit{for } (\alpha, \beta) \in g_\lambda. \tag{3.13}$$

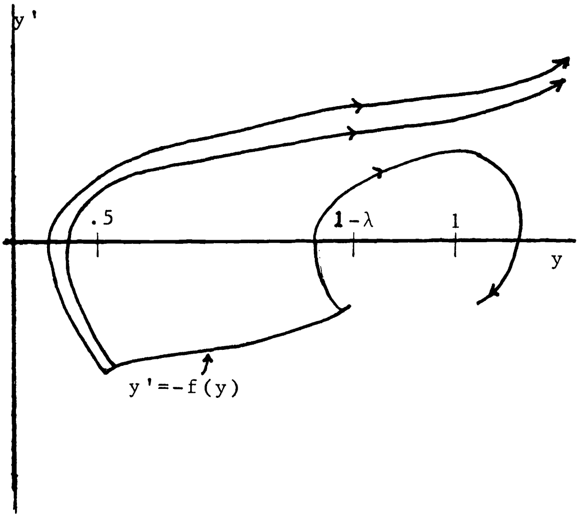


FIG. 7

The functional S plays a key role in proving the next pivotal result.

LEMMA 18. *Let $\frac{1}{2} \leq \alpha \leq 1-\lambda$ and $(y(0),(dy/dt)(0)) = (\alpha, -f(\alpha))$. If there is a value $\hat{t} > 0$ such that $S(\hat{t}) \geq 0$ and $y'' > 0$ for $0 < t \leq \hat{t}$, then $y'' > 0$ for all $t \in (0, \tilde{t})$.*

Proof. If $y''(t_0) = 0$ at some first $t_0 \in (\hat{t}, \tilde{t})$, then $y'''(t_0) \leq 0$ so that $y(t_0) \geq 1$. Thus, by (3.11), $S(t_0) \leq 0$. However, since $y'' > 0$ over $(\hat{t}, t_0)$, (3.12) and the condition $S(\hat{t}) \geq 0$ imply that $S(t_0) > 0$, a contradiction.

LEMMA 19. *If $\alpha - \frac{1}{2} \geq 0$ is sufficiently small and $(y(0),(dy/dt)(0)) = (\alpha, -f(\alpha))$, then there exists $\hat{t} = \hat{t}(\alpha) > 0$ such that $S(\hat{t}) > 0$, and $y'' > 0$ over $(0, \hat{t}]$.*

Proof. First we assume that $\alpha = \frac{1}{2}$ so that $y(0) = \frac{1}{2}, (dy/dt)(0) = -f(\frac{1}{2}) \leq 0$ and $(d^2y/dt^2)(0) = 0$. Then $y'''(0) > 0$ and $S(0) = 0$. Thus $y'' > 0$ and $S > 0$ on an interval $(0, \epsilon)$. This, continuity, and Lemma 18 prove the result.

Lemmas 17–19 will be used to help complete the first part of the proof of Theorem 2. For this we define the set $J = \{\hat{\alpha} \in (\frac{1}{2}, 1) |$ if $\frac{1}{2} < \alpha < \hat{\alpha}$ and $(y(0), y'(0), y''(0)) = (\alpha, -f(\alpha), 0)$, then there exists $\hat{t} > 0$ such that $y'' > 0$ for $0 < t \leq \hat{t}$ and $S(\hat{t}) > 0\}$. Continuity and Lemma 19 imply that J is open and nonempty.

Let $\alpha^* \equiv \sup J$. It follows from Lemma 17 that $\alpha^* < 1 - \lambda$. We consider the special solution of (3.10) for which $(y(0), (dy/dt)(0), (d^2y/dt^2)(0)) = (\alpha^*, -f(\alpha^*), 0)$.

We need to prove that

$$\lim_{t\to\infty} \left(y(t), \frac{dy}{dt}, \frac{d^2y}{dt^2} \right) = (\infty, \infty, 0). \tag{3.14}$$

First, since $y(0) = \alpha^* < 1$, then (3.9) implies that $y'' > 0$ for $t > 0$ at least until $y(t_1) = 1$ at some first $t_1 > 0$. Also, since $y(0) < 1$, $y'(0) \leq 0$ and $y'' > 0$ over $(0, t_1)$, it must be the case that $y'(t_1) > 0$. Thus we have $y(t_1) = 1, y'(t_1) > 0$, and $y''(t_1) > 0$. If there were a first $t_0 > t_1$ for which $y''(t_0) = 0$, then $y(t_0) > 1, y'''(t_0) < 0$, and $S(t_0) < 0$. It follows from (3.12) that $S < 0$ for $0 < t \leq t_0$. From this and continuity we see that $\alpha \notin J$ if $\alpha^* - \alpha \geq 0$ is sufficiently small, contradicting the definition of α^*. Thus it must be the case that $y'' > 0, y > 1, y' > y'(t_1) > 0,$ and $y''' < 0$ over (t_1, ∞). From this we conclude that $y(\infty) = \infty$ and $y''(\infty)$ must exist. If $y''(\infty) > 0$, then $S(\infty) = \infty$. Therefore, $S(\hat{t}) > 0$ for some $\hat{t} > 0$ and $y'' > 0$ over $(0, \hat{t}]$. But then continuity implies that $\alpha \in J$ if $\alpha - \alpha^* > 0$ is sufficiently small, again contradicting the definition of α^*. Therefore, $y''(\infty) = 0$. Finally, suppose that $y'(\infty) = K$ for some $K \in (0, \infty)$. Then, for large t it must be the case that $Kt < y < 2Kt$ and $y''' \le -1/(8K^2t^2)$. Then $y''(t) > \int_t^\infty (1/(8K^2u^2))du = 1/(8k^2t)$. One more integration shows that y' exceeds K at some finite t, a contradiction of the definition of K unless $K = \infty$. Thus we conclude that $\lim_{t\to\infty} y'(t) = \infty$.

To finish the proof of the first part of Theorem 2 we recall that $t = -x$ and conclude from the properties given above that $\lim_{x\to-\infty}(y(x), dy/dx, d^2y/dx^2) = (\infty, -\infty, 0)$.

The final step in proving Theorem 2 is to show that

$$\lim_{t\to\infty} \frac{y(t)}{(3^{1/3}t \log^{1/3}(t))} = 1 \tag{3.15}$$

and

$$\lim_{t\to\infty} \frac{y'(t)}{(3^{1/3} \log^{1/3}(t))} = 1. \tag{3.16}$$

At several points in the derivation of (3.15) and (3.16) we will make use of the following version of l'Hôpital's Rule.

L'Hôpital's Rule [2, pp. 384–385]. Suppose that f and g are differentiable and $g'(t) \neq 0$ on an open interval (d, ∞). Suppose that

$$\lim_{t\to\infty} f(t) = \lim_{t\to\infty} g(t) = 0 \quad \text{or that}$$
$$\lim_{t\to\infty} f(t) = \pm\infty \quad \text{and} \quad \lim_{t\to\infty} g(t) = \pm\infty.$$

If $\lim_{t\to\infty} f'(t)/g'(t)$ exists, then $\lim_{t\to\infty} f(t)/g(t) = \lim_{t\to\infty} f'(t)/g'(t)$.

We need one last technical result for the proof of (3.15) and (3.18), namely, the following.

LEMMA 20. *Let* $(y(0), (dy/dt)(0), (d^2y/dt^2)(0)) = (\alpha^*, -f(\alpha^*), 0)$. *Then*

(i) $\lim_{t\to\infty} S(t) = 0$;
(ii) $\lim_{t\to\infty} t/y = \lim_{t\to\infty} ty'' = \lim_{t\to\infty} yy'' = 0$;
(iii) $\lim_{t\to\infty} yy'y'' = 1$;
(iv) $\lim_{t\to\infty}((y')^3/\log(y)) = 3$;
(v) $\lim_{t\to\infty}(\log(y)/\log(t)) = 1$.

Proof. (i) Recall that $S = y'y'' - 1/y + 1/(2y^2)$ and $y(\infty) = \infty$. Thus we only need to prove that $y'y'' \longrightarrow 0$ as $t \longrightarrow \infty$. If $S(\hat{t}) > 0$ at some $\hat{t} > 0$, then continuity again implies that $\alpha^* \neq \sup J$. Therefore, $S(t) < 0$ on $(0, \infty)$, and it follows that $\lim_{t\to\infty} y'y'' \le 0$. However, $y' > 0$ and $y'' > 0$ for large $t > 0$. Thus $\lim_{t\to\infty} y'y'' = 0$, and (i) is proved.

(ii) From l'Hôpital's Rule we see that $\lim_{t\to\infty} t/y = \lim_{t\to\infty} 1/y' = 0$. An integration of (3.9) leads to

$$y''(t) = \int_t^\infty (y^{-2} - y^{-3})d\mu. \tag{3.17}$$

From (3.17) and an application of l'Hôpital's Rule we obtain

$$\lim_{t\to\infty} ty'' = \lim_{t\to\infty} \left(\int_t^\infty (y^{-2} - y^{-3})d\mu\right) \Big/ \left(\frac{1}{t}\right) = \lim_{t\to\infty} \left(\frac{t}{y}\right)^2 \left(1 - \frac{1}{y}\right) = 0.$$

Similarly, $\lim_{t\to\infty} yy'' = \lim_{t\to\infty}(1-(1/y))(1/y') = 0$.

(iii) Next, we integrate $S' = (y'')^2$ from t to ∞ and obtain

$$y'y'' = \frac{1}{y} - \frac{1}{2y^2} - \int_t^\infty (y'')^2 d\mu. \tag{3.18}$$

Multiplication of (3.18) by y gives

$$yy'y'' = 1 - \frac{1}{2y} - y\int_t^\infty (y'')^2 d\mu. \tag{3.19}$$

We need to show that the last term in (3.19) tends to zero as $t \longrightarrow \infty$. Again, by l'Hôpital's Rule, we obtain

$$\lim_{t\to\infty} y\int_t^\infty (y'')^2 d\mu = \lim_{t\to\infty} \frac{(yy'')^2}{y'} = 0.$$

(iv) From l'Hôpital's Rule, and (iii), it follows that

$$\lim_{t\to\infty} \frac{(y')^3}{\log(y)} = \lim_{t\to\infty} 3\, yy'y'' = 3.$$

(v) From (ii) and l'Hôpital's Rule, we obtain

$$\lim_{t\to\infty} \frac{\log(y)}{\log(t)} = \lim_{t\to\infty} \frac{ty'}{y} = \lim_{t\to\infty} \left(1 + \frac{ty''}{y'}\right) = 1.$$

We are now prepared to complete the derivation of (3.15) and (3.16). Multiply (3.18) by y', integrate from zero to t, and it follows that

$$\frac{(y')^3}{3} = \log(y) + \frac{1}{2y} - \int_0^t y'(\mu)\int_\mu^\infty (y''(s))^2 ds\, d\mu + W, \tag{3.20}$$

where $W = -[(f(\alpha^*))^3/3 + \log(\alpha^*) + 1/\alpha^*]$.

We divide (3.20) by $\log(t)$, use (v) of Lemma 20 and l'Hôpital's Rule to obtain

$$\lim_{t\to\infty} \frac{(y'(t))^3}{3\log(t)} = 1 + \lim_{t\to\infty}\left(-ty'\int_t^\infty (y''(s))^2 ds\right). \tag{3.21}$$

We need to show that the last term in (3.21) tends to zero as $t \to \infty$. First, since $y' > 0$ for large $t > 0$, then (3.21) implies that $y' \le 2[3^{1/3}\log^{1/3}(t)]$ for large $t > 0$. Thus

$$ty'(t)\int_t^\infty (y''(s))^2 ds \le 2t3^{1/3}\log^{1/3}(t)\int_t^\infty (y''(s))^2 ds. \tag{3.22}$$

From l'Hôpital's Rule we conclude that

$$\begin{aligned} &\lim_{t\to\infty} \int_t^\infty \frac{(y''(s))^2 ds}{(t^{-1}\log^{-1/3}(t))} \\ &= \lim_{t\to\infty} \frac{t^2(y''(t))^2\log^{1/3}(t)}{(1+3^{-1}\log^{-1}(t))}. \end{aligned} \tag{3.23}$$

It follows from parts (iii), (iv), and (v) of Lemma 20 that for large $t > 0$,

$$\lim_{t\to\infty} ty''(t)\log^{1/6}(t) = \lim_{t\to\infty}\left(\frac{t}{y}\right)(yy'y'')\left(\frac{\log(t)}{\log(y)}\right)^{1/6}\frac{\log^{1/6}(y)}{y'} = 0.$$

This and (3.23) imply that the right-hand side of (3.22) tends to zero as $t \longrightarrow \infty$. Thus, from (3.21) it follows that $\lim_{t\to\infty}(y'(t))^3/3\log(t) = 1$. Finally, this last limit, and l'Hôpital's Rule, leads to

$$\lim_{t\to\infty}\frac{y(t)}{(3^{1/3}t\log^{1/3}(t))}$$
$$= \lim_{t\to\infty}\frac{y'}{(3^{1/3}\log^{1/3}(t) + 3^{-2/3}\log^{-2/3}(t))} = 1.$$

This implies (3.16).

REFERENCES

[1] J. B. McLeod and J. Serrin, *The existence of similar solutions for some laminar boundary layer equations*, Arch. Rational Mech. Anal., 31 (1968), pp. 288–303.

[2] J. Stewart, *Calculus*, Brooks/Cole, Pacific Grove, CA, 1987.

[3] E. O. Tuck and L. W. Schwartz, *A numerical and asymptotic study of some third order differential equations relevant to draining and coating flows*, SIAM Rev., 32 (1990), pp. 453–469.

SIAM J. MATH. ANAL.
Vol. 24, No. 1, pp. 172-199, January 1993

011

ASYMPTOTIC BEHAVIOR OF SOLUTIONS OF SDE FOR RELAXATION OSCILLATIONS*

K. NARITA†

Abstract. A stochastic Liénard equation with a small parameter $\varepsilon > 0$ multiplying the highest derivative is formulated by a two-dimensional stochastic differential equation (SDE). Here fast and slow variables appear. In order to investigate the asymptotic behavior of the fast variable in such a system as $\varepsilon \to 0$, a stochastic process $X^\varepsilon(t)$ as a good approximation for all $t \geqq 0$ is derived by the methods of matched and composite expansions for relaxation oscillations. Then the limit of $X^\varepsilon(t)$ is identified, so that $X^\varepsilon(t)$ converges weakly as $\varepsilon \to 0$ to a solution of a one-dimensional stochastic differential equation. This yields the weak convergence of the slow variable as $\varepsilon \to 0$.

Key words. Liénard equation, stochastic differential equation, composite expansion, relaxation oscillation, relative compactness, weak convergence

AMS(MOS) subject classifications. 60H10, 60H20, 60J60, 34F05

1. Matched and composite asymptotic expansions. We consider the influence of random perturbations upon the Liénard oscillator:

$$\varepsilon \frac{d^2x}{dt^2} + \kappa \frac{dx}{dt} + g(x) = \delta \frac{dw}{dt}$$

with a family $\{\varepsilon, \kappa, \delta\}$ of positive constants, where d/dt denotes the symbolic derivative and dw/dt is the so-called white noise. Here and hereafter ε is a small parameter such that $0 < \varepsilon \ll 1$, and $g(x)$ is a scalar function on $x \in R^1 = (-\infty, \infty)$. Putting $y = \varepsilon(dx/dt) + \kappa x$, we get an equivalent system:

$$\varepsilon \frac{dx}{dt} = y - \kappa x, \qquad \frac{dy}{dt} = -g(x) + \delta \frac{dw}{dt}.$$

The object of interest is the asymptotic behavior of the solution

$$(x(t), y(t)), \quad 0 \leqq t \leqq T (T < \infty) \quad \text{as } \varepsilon \downarrow 0.$$

The above equation follows from the formal oscillator:

$$\frac{d^2x}{d\tau^2} + \nu\kappa \frac{dx}{d\tau} + g(x) = \sqrt{\nu}\, \delta \frac{db}{d\tau}$$

with a large parameter $\nu \gg 1$ and a white noise $db/d\tau$, under consideration that

$$\varepsilon = \frac{1}{\nu^2} \ll 1, \qquad t = \frac{\tau}{\nu}$$

and a Brownian motion process $w(t) = (1/\sqrt{\nu})b(\nu t)$.

When $\delta = 0$, this equation of the Liénard type is connected with explaining the *relaxation oscillations*, which is named after van der Pol [8].

While an equation takes the form

$$\frac{d^2x}{dt^2} + g(x) = \varepsilon f\left(x, \frac{dx}{dt}\right) + \sqrt{\varepsilon}\, \delta \frac{dw}{dt}, \qquad 0 < \varepsilon \ll 1$$

* Received by the editors June 10, 1991; accepted for publication (in revised form) May 5, 1992.

† Department of Mathematics, Faculty of Technology, Kanagawa University, Rokkakubashi Kanagawa-ku, Yokohama 221, Japan.

for which the *averaging principle* of Papanicolaou [5] applies, our oscillator is of the type

$$\varepsilon\frac{d^2x}{dt^2}=f\left(x,\frac{dx}{dt}\right)+\delta\frac{dw}{dt},\qquad 0<\varepsilon\ll 1.$$

So we adopt another method.

Let $(\Omega, \mathbf{F}, P)$ be a probability space with an increasing family $\{\mathbf{F}_t; t\geqq 0\}$ of sub-σ-algebras of $\mathbf{F}$, and let $w(t)$ be a one-dimensional Brownian motion process adapted to $\mathbf{F}_t$. Then as the response of the above oscillator we take the solution of the following two-dimensional stochastic differential equation:

$$\begin{aligned}\varepsilon\, dx(t)&=[y(t)-\kappa x(t)]\,dt,\\ dy(t)&=-g(x(t))\,dt+\delta\,dw(t).\end{aligned}\tag{1.1}$$

Let $\phi=(\xi, \eta)$ be a two-dimensional random vector independent of the two-dimensional Brownian motion process. Then we consider (1.1) under the following initial condition:

$$x(0)=\xi,\qquad y(0)=\eta.\tag{1.2}$$

For $\varepsilon=0$ we obtain the *reduced system*:

$$0=y_0-\kappa x_0,\qquad dy_0=-g(x_0)\,dt+\delta\,dw(t).$$

Namely,

$$\begin{aligned}dx_0(t)&=-\frac{1}{\kappa}g(x_0(t))\,dt+\frac{\delta}{\kappa}\,dw(t),\\ y_0(t)&=\kappa x_0(t)\end{aligned}\tag{1.3}$$

with the initial state $(x_0(0), y_0(0))=(C_0, \kappa C_0)$, where the value C_0 is still undetermined. Since the derivative of $x(t)$ is of order $O(1/\varepsilon)$ formally, $x(t)$ is rapidly varying until the line $y=\kappa x$ is reached. During this short period of time, $y(t)$ seems to remain almost unchanged. Thus it is plausible that in the limit $\varepsilon\to 0$ the solution jumps at $t=0$ from the initial state (ξ, η) to the state$(\eta/\kappa, \eta)$. This enables us to determine C_0 in approximate solution (1.3) valid for $t>0$: $C_0=\eta/\kappa$. Obviously, this approximate solution $x_0(t)$ with the initial state $x_0(0)=C_0=\eta/\kappa$ does not satisfy the initial condition (1.2). In order to obtain a good approximation in both the time intervals such that

$$\text{near } t=0,\quad \text{away from } t=0,$$

we adopt the methods of *matched and composite expansions* in Grasman [1, Chap. 1] and Nayfeh [3, Chap. 4].

In § 1 we derive a good approximation for (1.1) in *all time intervals.* In § 2 we identify the limit, so that such an approximation converges weakly as $\varepsilon\to 0$ to a diffusion process governed by (1.3). In §§ 3 and 4 we prepare some estimates for the moment of the solution of (1.1), and in § 5 we give the proof of the identification of the limit. In the following, without being concerned for the moment with mathematical rigor, we present some preliminary considerations.

Since ε is small, we naively look for a *regular perturbation solution* (Taylor series in ε) of the form

$$\begin{aligned}x(t;\varepsilon)&=x_0(t)+\varepsilon x_1(t)+\cdots,\\ y(t;\varepsilon)&=y_0(t)+\varepsilon y_1(t)+\cdots.\end{aligned}\tag{1.4}$$

This is called the *outer expansion*. Substituting (1.4) into (1.1) and systematically equating the coefficients of powers of ε we get the following equations:

$$\begin{aligned} 0 &= y_0 - \kappa x_0, \\ dy_0 &= -g(x_0)\,dt + \delta\,dw(t), \end{aligned} \tag{1.5}$$

$$\begin{aligned} dx_{n-1} &= [y_n - \kappa x_n]\,dt, \\ dy_n &= -g_n(x_0, x_1, \dots, x_n)\,dt, \qquad n = 1, 2, \dots . \end{aligned} \tag{1.6}$$

Here we assume that

$$g(x_0 + \varepsilon x_1 + \varepsilon^2 x_2 + \cdots) = \sum_{n=0}^{\infty} g_n(x_0, x_1, x_2, \dots, x_n)\varepsilon^n$$

with $g_0(x_0) = g(x_0)$ and

$$g_n(x_0, x_1, x_2, \dots, x_n) = \sum_{\substack{n_1+n_2+\cdots+n_k=n \\ 1\leqq k\leqq n}} \frac{x_{n_1}\cdot x_{n_2}\cdots x_{n_k}}{k!}\left[\frac{d^k g(s)}{ds^k}\right]_{s=x_0}. \tag{1.7}$$

Equation (1.5) is just the same with the reduced system (1.3). We impose the initial condition on (1.5) as follows:

$$(x_0(0), y_0(0)) = (C_0, \kappa C_0).$$

Next we introduce the time scale:

$$\tau = \frac{t}{\varepsilon}.$$

Let $(x(t), y(t))$ be the solution of (1.1) with the initial state (1.2), and set

$$x(\tau;\,\varepsilon) = x(\tau\varepsilon), \qquad y(\tau;\,\varepsilon) = y(\tau\varepsilon).$$

Then $x(\tau;\,\varepsilon)$ and $y(\tau;\,\varepsilon)$ satisfy the following equations:

$$\begin{aligned} dx(\tau;\,\varepsilon) &= [y(\tau;\,\varepsilon) - \kappa x(\tau;\,\varepsilon)]\,d\tau, \\ dy(\tau;\,\varepsilon) &= -\varepsilon g(x(\tau;\,\varepsilon))\,d\tau + \sqrt{\varepsilon}\,\delta\,d\tilde{w}(\tau), \\ x(0;\,\varepsilon) &= \xi, \qquad y(0;\,\varepsilon) = \eta, \end{aligned} \tag{1.8}$$

where $\tilde{w}(\tau)$ is a Brownian motion process defined by $\tilde{w}(\tau) = (1/\sqrt{\varepsilon})w(\tau\varepsilon)$. It is assumed that the solution of (1.8) can be expanded in a power series of $\sqrt{\varepsilon}$:

$$\begin{aligned} x(\tau;\,\varepsilon) &= u_0(\tau) + \sqrt{\varepsilon}\,u_1(\tau) + \varepsilon u_2(\tau) + \cdots, \\ y(\tau;\,\varepsilon) &= v_0(\tau) + \sqrt{\varepsilon}\,v_1(\tau) + \varepsilon v_2(\tau) + \cdots . \end{aligned} \tag{1.9}$$

This is called the *inner expansion*. Substituting (1.9) into (1.8) and equating the coefficients of powers of $\sqrt{\varepsilon}$, we obtain the following equations:

$$\begin{aligned} &\frac{du_0}{d\tau} = v_0 - \kappa u_0, \qquad u_0(0) = \xi, \\ &v_0(\tau) = \eta, \end{aligned} \tag{1.10}$$

$$\begin{aligned} &\frac{du_1}{d\tau} = v_1 - \kappa u_1, \qquad u_1(0) = 0, \\ &v_1(\tau) = \delta\tilde{w}(\tau), \qquad v_1(0) = 0, \end{aligned} \tag{1.11}$$

$$(1.12)\qquad \begin{aligned} &\frac{du_n}{d\tau}=v_n-\kappa u_n, \qquad u_n(0)=0,\\ &\frac{dv_n}{d\tau}=-g_{n-2}(u_0,u_1,\ldots,u_{n-2}), \qquad v_n(0)=0,\\ &n=2,3,\ldots, \end{aligned}$$

where $g(u_0+\sqrt{\varepsilon}\,u_1+\varepsilon u_2+\cdots)$ is assumed to be expanded as

$$g(u_0+\sqrt{\varepsilon}\,u_1+\varepsilon u_2+\cdots)=\sum_{n=0}^{\infty} g_n(u_0,u_1,u_2,\ldots,u_n)(\sqrt{\varepsilon})^n$$

with $g_0(u_0)=g(u_0)$ and $g_n(u_0,u_1,u_2,\ldots,u_n)$ as in (1.7). The solution of (1.10) is given by

$$(1.13)\qquad u_0(\tau)=\xi\exp[-\kappa\tau]+\frac{\eta}{\kappa}[1-\exp[-\kappa\tau]], \qquad v_0(\tau)=\eta.$$

We must determine the value $x_0(0)=C_0$ in the outer expansion (1.4) and (1.5) so that both expansions (1.4) and (1.9) should have a limited time interval over which they are valid. Since τ must be large and t small, we substitute $t=\varepsilon\tau$ in (1.4) and require the identical asymptotic behavior of both expansions for τ large. Namely, by (1.4), we have

$$\begin{aligned} x&=x_0(\tau\varepsilon)+\varepsilon x_1(\tau\varepsilon)+\cdots,\\ y&=y_0(\tau\varepsilon)+\varepsilon y_1(\tau\varepsilon)+\cdots, \end{aligned}$$

and so

$$\begin{aligned} x&=\left[C_0-\frac{1}{\kappa}\int_0^{\tau\varepsilon} g(x_0(s))\,ds+\frac{\delta}{\kappa}\,w(\tau\varepsilon)\right]+\varepsilon x_1(\tau\varepsilon)+\cdots,\\ y&=\left[\kappa C_0-\int_0^{\tau\varepsilon} g(x_0(s))\,ds+\delta w(\tau\varepsilon)\right]+\varepsilon y_1(\tau\varepsilon)+\cdots. \end{aligned}$$

On the other hand, for large τ,

$$(1.14)\qquad u_0\approx\frac{\eta}{\kappa}, \qquad v_0\approx\eta.$$

Thus it is necessary that

$$C_0=\frac{\eta}{\kappa}.$$

Add the two expansions (1.4) and (1.9) and subtract the matching terms (1.14). Then we get the following expansion, which may be valid in both time intervals:

$$(1.15)\qquad \begin{aligned} x(t;\varepsilon)&=\left[x_0(t)+u_0\left(\frac{t}{\varepsilon}\right)-\frac{\eta}{\kappa}\right]+O(\varepsilon),\\ y(t;\varepsilon)&=\left[y_0(t)+v_0\left(\frac{t}{\varepsilon}\right)-\eta\right]+O(\varepsilon). \end{aligned}$$

This corresponds to the so-called *composite expansion.*

Let $(x^\varepsilon(t), y^\varepsilon(t))$ be the solution of (1.1) with the initial state (1.2). Then, being motivated by the formal expansion (1.15), we take account of the following processes:

$$X^\varepsilon(t)=x^\varepsilon(t)-\left[u_0\left(\frac{t}{\varepsilon}\right)-\frac{\eta}{\kappa}\right], \qquad Y^\varepsilon(t)=y^\varepsilon(t)-\kappa X^\varepsilon(t).$$

Then, in § 2 we give a theorem on the weak convergence of $X^\varepsilon(t)$ and $y^\varepsilon(t)$ as $\varepsilon \to 0$ to the diffusion processes $x_0(t)$ and $y_0(t)$ governed by SDE (1.3), respectively. The proof of the theorem is given in § 5 by using the following estimates:

Section 3: uniform boundedness for the moment of $(x^\varepsilon(t), y^\varepsilon(t))$ with respect to ε;

Section 4: relative compactness for $X^\varepsilon(t)$ and $y^\varepsilon(t)$;

Section 6: appendix (a priori bounds of the solution of the Langevin equation with a drift multiplied by $1/\varepsilon$).

2. Theorem on weak convergence for $\varepsilon \to 0$. Let $(x^\varepsilon(t), y^\varepsilon(t))$ be the solution of (1.1) with the initial state (1.2). Then we define a pair $(X^\varepsilon(t), Y^\varepsilon(t))$ of random processes by

$$X^\varepsilon(t)=x^\varepsilon(t)-\left[u_0\left(\frac{t}{\varepsilon}\right)-\frac{\eta}{\kappa}\right], \tag{2.1}$$
$$Y^\varepsilon(t)=y^\varepsilon(t)-\kappa X^\varepsilon(t),$$

where $u_0(\tau)$ is given by (1.13). We notice that

$$X^\varepsilon(t)=x^\varepsilon(t)-\left(\xi-\frac{\eta}{\kappa}\right)\exp\left[-\frac{\kappa}{\varepsilon}t\right], \qquad X^\varepsilon(0)=\frac{\eta}{\kappa}, \tag{2.1)'$$
$$Y^\varepsilon(t)=[y^\varepsilon(t)-\kappa x^\varepsilon(t)]-(\eta-\kappa\xi)\exp\left[-\frac{\kappa}{\varepsilon}t\right], \qquad Y^\varepsilon(0)=0.$$

We shall need the following definition and assumptions.

DEFINITION 2.1. For the continuous function $g(x)$, set $G(x)=\int_0^x g(s)\,ds$ and assume that there exists a constant $\beta \geqq 0$ such that $G(x)+\beta \geqq 0$ for all $x \in R^1$. Then, for each $0<\varepsilon \leqq 1$ and $(x, y) \in R^2=R^1 \times R^1$, we define the functions $V_\varepsilon(x, y)$ and $V(x, y)$ by

$$V_\varepsilon(x, y)=\varepsilon[G(x)+\beta]+\frac{y^2}{2}, \qquad V(x, y)=[G(x)+\beta]+\frac{y^2}{2}.$$

Assumption 2.1. The scalar function $g(x)$ is once continuously differentiable with respect to $x \in R^1$, and there exists a constant $c>0$ such that

$$|g(x)|+|g'(x)| \leqq c(1+|x|^p) \quad \text{for all } x \in R^1$$

with an integer $p \geqq 1$.

Assumption 2.2. The function $g(x)$ satisfies the following conditions:

(i) There exists a constant $\alpha>0$ such that

$$-xg(x) \leqq \alpha \quad \text{for all } x \in R^1;$$

(ii) $G(x) \to \infty$ as $|x| \to \infty$, where $G(x)$ is as in Definition 2.1;

(iii) Let $\beta \geqq 0$ be a constant such that $G(x)+\beta \geqq 0$ for all $x \in R^1$. Then there exists a constant $l>0$ such that

$$x^2+y^2 \leqq lV(x, y) \quad \text{for all } (x, y) \in R^2,$$

where $V(x, y)$ is as in Definition 2.1.

Assumption 2.3. Let $\phi=(\xi, \eta)$ be a two-dimensional random vector independent of the two-dimensional Brownian motion process, such that

$$E[V(\phi)^{mq}]<\infty \quad \text{for an integer } m \geqq 2,\ q=\max\{2p, 2\},$$

where p is as in Assumption 2.1 and V is as in Definition 2.1.

Example 2.1. Consider the following functions.

(i) $g(x)=\sum_{k=0}^{n}(2k+2)\alpha_{2k+2}x^{2k+1}$ with a family $\{\alpha_{2k+2}\}$ of positive constants;

(ii) $g(x)=rx^3-sx$ with constants $r>0$ and $s>0$.

In case (i), Assumptions 2.1 and 2.2 hold for $g(x)$ with the following choice of constants:

$$c=\sum_{k=0}^{n}(2k+2)^2\alpha_{2k+2}, \qquad p=2n+1,$$

$$\alpha>0, \quad \beta\geqq 0, \quad l\geqq \max\{2, \alpha_2^{-1}\}.$$

In case (ii), Assumptions 2.1 and 2.2 hold for $g(x)$ with the following choice of constants:

$$c=4r+2s, \qquad p=3,$$

$$\alpha=\frac{s^2}{4r^2}, \qquad \beta>\frac{s^2}{4r^2}, \qquad l\geqq \max\left\{2, \left(\sqrt{r\beta}-\frac{s}{2}\right)^{-1}\right\}.$$

Hereafter, $C([0, T]; R^1)$ is the space of continuous paths ω: $\to R^1$.

THEOREM 2.1. *Suppose that Assumptions* 2.1, 2.2, *and* 2.3 *hold. For the solution* $(x^\varepsilon(t), y^\varepsilon(t))$ *of* SDE (1.1) *with the initial state* (1.2), *let* $(X^\varepsilon(t), Y^\varepsilon(t))$ *be the process defined by* (2.1). *Then* $X^\varepsilon(t)$, $t\geqq 0$, *with the initial state* $X^\varepsilon(0)=\eta/\kappa$, *converges weakly in* $C([0, T]; R^1)$, $T<\infty$, *but arbitrary, as* $\varepsilon\to 0$ *to the solution* $X(t)$ *of the following one-dimensional* SDE:

$$dX(t)=-\frac{1}{\kappa}g(X(t))\,dt+\frac{\delta}{\kappa}d\tilde{w}(t), \qquad X(0)=\frac{\eta}{\kappa}, \tag{2.2}$$

where $\tilde{w}(t)$ *is a one-dimensional Brownian motion process. Moreover,* $E[\sup_{0\leqq t\leqq T}|Y^\varepsilon(t)|^{2m}]\to 0$ *as* $\varepsilon\to 0$ *for every* $T<\infty$ *with the same exponent m as in Assumption* 2.3. *In particular,* $y^\varepsilon(t)$, $t\geqq 0$, *with the initial state* $y^\varepsilon(0)=\eta$, *converges weakly in* $C([0, T]; R^1)$, $T<\infty$, *but arbitrary, as* $\varepsilon\to 0$ *to the process* $\kappa X(t)$.

3. Uniform boundedness for the moment. Let $z^\varepsilon(t)=(x^\varepsilon(t), y^\varepsilon(t))$ be the solution of SDE (1.1) with the initial state (1.2): $z^\varepsilon(0)=\phi$. Hereafter, ϕ is a two-dimensional random vector independent of the two-dimensional Brownian motion process, and V_ε and V are the functions given by Definition 2.1. Then, for the proof of Theorem 2.1 we estimate the moment of $z^\varepsilon(t)$ from above uniformly in ε.

LEMMA 3.1. *Suppose that* $g(x)$ *satisfies the conditions* (i) *and* (ii) *of Assumption* 2.2. *Suppose that* ϕ *satisfies the following moment condition*:

$$E[V(\phi)^m]<\infty \quad \textit{for an integer } m\geqq 1.$$

Then there exists a pathwise unique solution $z^\varepsilon(t)=(x^\varepsilon(t), y^\varepsilon(t))$ *of* (1.1) *with the initial state* $z^\varepsilon(0)=\phi$. *For each* $0<\varepsilon\leqq 1$, *set* $U_\varepsilon(t)=V_\varepsilon(z^\varepsilon(t))$. *Then*

$$E[(1+U_\varepsilon(t))^m]\leqq E[(1+V(\phi))^m]\exp[c_m t], \tag{3.1}$$

where $c_m=m[\alpha\kappa+\delta^2(m-\frac{1}{2})]$.

Proof. Denote by L^ε the differential generator associated with (1.1). Namely, for $(x, y)\in R^2$,

$$L^\varepsilon=\frac{1}{\varepsilon}(y-\kappa x)\frac{\partial}{\partial x}-g(x)\frac{\partial}{\partial y}+\frac{1}{2}\delta^2\frac{\partial^2}{\partial y^2}.$$

Then, by condition (i) of Assumption 2.2, V_ε satisfies

$$L^\varepsilon V_\varepsilon(x, y)=-\kappa x g(x)+\tfrac{1}{2}\delta^2\leqq \alpha\kappa+\tfrac{1}{2}\delta^2 \quad \text{for all } (x, y)\in R^2.$$

By the condition (ii) of Assumption 2.2, V_ε is radially unbounded: $V_\varepsilon(x, y) \to \infty$ as $(x^2+y^2)^{1/2} \to \infty$. Notice that $E[V_\varepsilon(\phi)] \leqq E[V(\phi)] < \infty$. So, according to nonexplosion criteria in Hasminskii [2] and Narita [4], any solution of SDE (1.1) with the initial state (1.2) such that $E[V_\varepsilon(\phi)] < \infty$ cannot explode, which implies the pathwise uniqueness for the solution. Moreover,

$$L^\varepsilon[(1+V_\varepsilon)^m] \leqq c_m(1+V_\varepsilon)^m \tag{3.2}$$

with the constant $c_m = m[\alpha\kappa + \delta^2(m-\frac{1}{2})]$. Apply the Ito formula to $(1+U_\varepsilon(t))^m$ and take expectation. Then, by (3.2) and the Gronwall lemma, we get (3.1). Hence the proof is complete.

LEMMA 3.2. *Under the same assumptions as in Lemma* 3.1, *suppose that the moment condition on ϕ is replaced by the following condition*:

$$E[V(\phi)^{2m}] < \infty \quad \textit{for an integer } m \geqq 1.$$

For each $0 < \varepsilon \leqq 1$, *set* $U_\varepsilon(t) = V_\varepsilon(z^\varepsilon(t))$. *Then*

$$E\left[\sup_{0 \leqq t \leqq T} (1+U_\varepsilon(t))^m\right] \leqq A_m(T) \exp[c_m T] \tag{3.3}$$

for every $T < \infty$, *where*

$$A_m(T) = E[(1+V(\phi))^m] + B_m \exp\left[\frac{c_{2m}T}{2}\right],$$

$$B_m = 2\delta m(2c_{2m}^{-1})^{1/2}(E[(1+V(\phi))^{2m}])^{1/2},$$

and c_{2m} *is the constant obtained by* c_m *in* (3.1) *with m replaced by* 2*m*.

Proof. For each $T < \infty$, set $S_\varepsilon(T) = \sup_{0 \leqq t \leqq T}(1+U_\varepsilon(t))^m$, where $U_\varepsilon(t) = V_\varepsilon(x^\varepsilon(t), y^\varepsilon(t))$. Then (3.2) yields

$$S_\varepsilon(T) \leqq S_\varepsilon(0) + c_m \int_0^T S_\varepsilon(t)\,dt + \sup_{0 \leqq t \leqq T} M_\varepsilon(t), \tag{3.4}$$

where $M_\varepsilon(t) = \int_0^t m(1+U_\varepsilon(s))^{m-1}\delta y^\varepsilon(s)\,dw(s)$. Put $e_N = \inf\{t; |z^\varepsilon(t)| \geqq N\}$. Then Lemma 3.1 implies that $e_N \to \infty$ with probability 1 as $N \to \infty$. Since $y^2 \leqq 2V_\varepsilon(z)$ for $z = (x, y) \in R^2$, $M_\varepsilon(t)$ satisfies

$$E[M_\varepsilon(t \wedge e_N)^2] \leqq 2\delta^2 m^2 \int_0^t E[(1+U_\varepsilon(s \wedge e_N))^{2m}]\,ds,$$

where $a \wedge b$ is the smaller of a and b. Notice that $E[V(\phi)^{2m}] < \infty$, so that the estimate (3.1) of Lemma 3.1 holds with m replaced by $2m$. Thus

$$E[M_\varepsilon(t)^2] = \lim_{N \to \infty} E[M_\varepsilon(t \wedge e_N)^2] < \infty.$$

Namely, $\{M_\varepsilon(t);\ t \geqq 0\}$ is a square integrable martingale satisfying

$$E[M_\varepsilon(t)^2] \leqq 2\delta^2 m^2 E[(1+V(\phi))^{2m}] \int_0^t \exp[c_{2m}s]\,ds,$$

where c_{2m} is the constant obtained by c_m in (3.1) with m replaced by $2m$. Thus the martingale inequality yields

$$\begin{aligned} E\left[\sup_{0 \leqq t \leqq T} |M_\varepsilon(t)|\right] &\leqq \left(E\left[\left\{\sup_{0 \leqq t \leqq T} M_\varepsilon(t)\right\}^2\right]\right)^{1/2} \\ &\leqq (4E[M_\varepsilon(T)^2])^{1/2} \\ &\leqq 2(2\delta^2 m^2 E[(1+V(\phi))^{2m}])^{1/2} \\ &\quad \times (c_{2m}^{-1}[\exp[c_{2m}T]-1])^{1/2}. \end{aligned}$$

Take expectation on (3.4), so that

$$E[S_\varepsilon(T)] \leqq E[S_\varepsilon(0)] + c_m \int_0^T E[S_\varepsilon(t)]\,dt + E\left[\sup_{0\leqq t\leqq T} |M_\varepsilon(t)|\right].$$

Further, notice that

$$E[S_\varepsilon(0)] \leqq E[(1+V(\phi))^m] \leqq (E[(1+V(\phi))^{2m}])^{1/2} < \infty$$

and that $(\exp[xT]-1)^{1/2} \leqq \exp[xT/2]$ for $x \geqq 0$. Then by the Gronwall lemma we get (3.3). Hence the proof is complete.

LEMMA 3.3. *Let the same assumptions as in Lemma* 3.2 *hold. Then*

$$E\left[\sup_{0\leqq t\leqq T} |y^\varepsilon(t)|^{2m}\right] \leqq 2^m A_m(T) \exp[c_m T] \quad \textit{for every } T<\infty, \tag{3.5}$$

where c_m and $A_m(T)$ are the same constants as in (3.1) *and* (3.3), *respectively.*

Proof. By Definition 2.1, since $y^2/2 \leqq V_\varepsilon(x, y)$ for $(x, y) \in R^2$, we see that $(\frac{1}{2})^m y^{2m} \leqq (1+V_\varepsilon(x, y))^m$. Thus (3.5) follows from (3.3) of Lemma 3.2, and hence the proof is complete.

LEMMA 3.4. *Suppose that $g(x)$ satisfies all conditions of Assumption* 2.2 *and that ϕ satisfies the following moment condition*:

$$E[V(\phi)^{2m}] < \infty \quad \textit{for an integer } m \geqq 1.$$

Then

$$E\left[\sup_{0\leqq t\leqq T} |x^\varepsilon(t)|^{2m}\right] \leqq K_m(T) \quad \textit{for every } T<\infty, \tag{3.6}$$

where

$$K_m(T) = 2^{2m-1}[l^m E[V(\phi)^m] + \kappa^{-2m} 2^m A_m(T) \exp[c_m T]]$$

with the same constants c_m and $A_m(T)$ as in (3.1) *and* (3.3), *respectively.*

Proof. By SDE (1.1), $x^\varepsilon(t)$ satisfies the linear ordinary differential equation of first order:

$$\frac{d}{dt} x^\varepsilon(t) = \frac{1}{\varepsilon}[y^\varepsilon(t) - \kappa x^\varepsilon(t)], \qquad x^\varepsilon(0) = \xi.$$

Namely,

$$x^\varepsilon(t) = \xi \exp\left[\frac{-\kappa t}{\varepsilon}\right] + \frac{1}{\varepsilon} \exp\left[\frac{-\kappa t}{\varepsilon}\right] \int_0^t y^\varepsilon(s) \exp\left[\frac{\kappa s}{\varepsilon}\right] ds,$$

and so, for $0 \leqq t \leqq T$,

$$|x^\varepsilon(t)| \leqq |\xi| \exp\left[\frac{-\kappa t}{\varepsilon}\right] + \sup_{0\leqq t\leqq T} |y^\varepsilon(t)| \frac{1}{\kappa}\left[1 - \exp\left[\frac{-\kappa t}{\varepsilon}\right]\right].$$

Use the following inequalities:

$$(a+b)^{2m} \leqq 2^{2m-1}(a^{2m} + b^{2m}) \quad \text{for } a \geqq 0 \quad \text{and} \quad b \geqq 0,$$

$$0 \leqq \exp[-x] \leqq 1 \quad \text{and} \quad 0 \leqq 1 - \exp[-x] \leqq 1 \quad \text{for } x \geqq 0.$$

Then, for each $T > 0$,

$$\sup_{0\leqq t\leqq T} |x^\varepsilon(t)|^{2m} \leqq 2^{2m-1}\left[|\xi|^{2m} + \kappa^{-2m} \sup_{0\leqq t\leqq T} |y^\varepsilon(t)|^{2m}\right]. \tag{3.7}$$

By condition (iii) of Assumption 2.2, since $\phi=(\xi, \eta)$ satisfies

$$|\xi|^{2m} \leqq |\phi|^{2m} \leqq [lV(\phi)]^m,$$

the moment condition on ϕ implies

$$E[|\xi|^{2m}] \leqq l^m E[V(\phi)^m] \leqq l^m (E[V(\phi)^{2m}])^{1/2} < \infty.$$

Therefore, taking expectation on (3.7), by (3.5) of Lemma 3.3 we obtain (3.6). Hence the proof is complete.

4. Relative compactness for $X^\varepsilon(t)$ $y^\varepsilon(t)$. Hereafter let $(x^\varepsilon(t), y^\varepsilon(t))$ denote the solution of (1.1) with the initial state (1.2), i.e., $(x^\varepsilon(0), y^\varepsilon(0)) = \phi = (\xi, \eta)$, where ϕ is a two-dimensional random vector independent of the two-dimensional Brownian motion process. Throughout this section, let V be the function given by Definition 2.1.

LEMMA 4.1. *Suppose that there exists a constant $c>0$ such that*

$$|g(x)| \leqq c(1+|x|^p) \quad \textit{for all } x \in R^1 \textit{ with an integer } p \geqq 1,$$

and suppose that Assumption 2.2 *holds. Let ϕ be a random vector such that*

$$E[V(\phi)^{2q}] < \infty, \qquad q = \max\{2p, 2\},$$

where p is as in the hypothesis. Let $T<\infty$ be arbitrary and fixed. Then

$$\sup_{0<\varepsilon\leqq 1} E[|y^\varepsilon(t)-y^\varepsilon(s)|^4] \leqq D(t-s)^2 \quad \textit{for } 0 \leqq s \leqq t \leqq T \tag{4.1}$$

with a constant $D>0$ independent of ε.

Proof. According to the second equation of SDE (1.1) and the Schwarz inequality, $y^\varepsilon(t)$ satisfies

$$(y^\varepsilon(t)-y^\varepsilon(s))^4 \leqq 2^3\left[(t-s)^3 \int_s^t |g(x^\varepsilon(u))|^4\,du + \left(\int_s^t \delta\,dw(u)\right)^4\right] \tag{4.2}$$

for $0 \leqq s \leqq t \leqq T.$

The growth restriction on $g(x)$ implies

$$|g(x)|^4 \leqq c^4 2^3 (1+|x|^{4p}) \quad \text{for all } x \in R^1.$$

Put $m=2p$. Then, by virtue of the moment condition on ϕ we notice that

$$E[V(\phi)^{2m}] \leqq (E[V(\phi)^{2q}])^{2p/q} < \infty \quad \text{with } q = \max\{2p, 2\}.$$

So the estimate (3.6) of Lemma 3.4 holds with $m=2p$. Namely,

$$E\left[\sup_{0\leqq t\leqq T} |x^\varepsilon(t)|^{4p}\right] \leqq K_{2p}(T),$$

where $K_{2p}(T)$ is the constant obtained by $K_m(T)$ in (3.6) with $m=2p$.

On the other hand,

$$E\left[\left(\int_s^t \delta\,dw(u)\right)^4\right] \leqq \delta^4 36(t-s)^2 \quad \text{for } 0 \leqq s \leqq t \leqq T.$$

Taking expectation on (4.2), we get (4.1). Hence the proof is complete.

LEMMA 4.2. *Under the same assumptions as in Lemma* 4.1, *suppose that the moment condition on ϕ is replaced by the following condition*:

$$E[V(\phi)^{mq}] < \infty \quad \textit{for an integer } m \geqq 2, \qquad q = \max\{2p, 2\},$$

where p is as in the hypothesis of Lemma 4.1. *Let* $Y^\varepsilon(t)$ *be the process defined by* (2.1). *Then*

$$E\left[\sup_{0\leqq t\leqq T}|Y^\varepsilon(t)|^{2m}\right]\leqq F_m(T)(\varepsilon+\sqrt{\varepsilon})\quad \textit{for every } T<\infty, \tag{4.3}$$

where $F_m(T)$ *is a positive constant independent of* ε *such that* $F_m(T)$ *is a continuous and increasing function of T. Moreover, for every* $T<\infty$,

$$\sup_{0<\varepsilon\leqq 1} E[|Y^\varepsilon(t)-Y^\varepsilon(s)|^4]\leqq \tilde{D}(t-s)^2 \quad \textit{for } 0\leqq s\leqq t\leqq T \tag*{(4.3)$'$}$$

with a constant $\tilde{D}>0$ *independent of* ε.

Proof. First we show (4.3). According to (2.1)$'$, $Y^\varepsilon(t)$ has the following form:

$$Y^\varepsilon(t)=\Delta^\varepsilon(t)-(\eta-\kappa\xi)\exp\left[\frac{-\kappa t}{\varepsilon}\right],$$

$$\Delta^\varepsilon(t)=y^\varepsilon(t)-\kappa x^\varepsilon(t).$$

Since $(x^\varepsilon(t), y^\varepsilon(t))$ satisfies SDE (1.1) with the initial state $(x^\varepsilon(0), y^\varepsilon(0))=(\xi,\eta)=\phi$, $\Delta^\varepsilon(t)$ satisfies the following linear stochastic differential equation:

$$d\Delta^\varepsilon(t)=\left[-\frac{\kappa}{\varepsilon}\Delta^\varepsilon(t)-g(x^\varepsilon(t))\right]dt+\delta\, dw(t),$$

$$\Delta^\varepsilon(0)=\eta-\kappa\xi.$$

Therefore, the solution $\Delta^\varepsilon(t)$ can be written in the form

$$\Delta^\varepsilon(t)=I_1^\varepsilon(t)-I_2^\varepsilon(t)+I_3^\varepsilon(t),$$

where

$$I_1^\varepsilon(t)=\Delta^\varepsilon(0)\exp\left[\frac{-\kappa t}{\varepsilon}\right]=(\eta-\kappa\xi)\exp\left[\frac{-\kappa t}{\varepsilon}\right],$$

$$I_2^\varepsilon(t)=\exp\left[\frac{-\kappa t}{\varepsilon}\right]\int_0^t \exp\left[\frac{\kappa s}{\varepsilon}\right]g(x^\varepsilon(s))\,ds,$$

$$I_3^\varepsilon(t)=\exp\left[\frac{-\kappa t}{\varepsilon}\right]\int_0^t \exp\left[\frac{\kappa s}{\varepsilon}\right]\delta\, dw(s).$$

Accordingly, we have

$$Y^\varepsilon(t)=-I_2^\varepsilon(t)+I_3^\varepsilon(t)\quad \text{for all } t\geqq 0.$$

In the following, let $m\geqq 1$ be any integer and fixed. Then, for each $T<\infty$,

$$\sup_{0\leqq t\leqq T}|Y^\varepsilon(t)|^{2m}\leqq 2^{2m-1}\left[\sup_{0\leqq t\leqq T}|I_2^\varepsilon(t)|^{2m}+\sup_{0\leqq t\leqq T}|I_3^\varepsilon(t)|^{2m}\right]. \tag{4.4}$$

First we evaluate $|I_2^\varepsilon(t)|^{2m}$. Consider the growth restriction on $g(x)$, so that

$$|g(x)|^{2m}\leqq c^{2m}2^{2m-1}(1+|x|^{2mp})\quad \text{for all } x\in R^1.$$

Then the Schwarz inequality yields

$$\begin{aligned}|I_2^\varepsilon(t)|^{2m} &\leqq t^{2m-1}\exp\left[\frac{-2m\kappa t}{\varepsilon}\right]\\ &\quad\times\int_0^t \exp\left[\frac{2m\kappa s}{\varepsilon}\right]c^{2m}2^{2m-1}(1+|x^\varepsilon(s)|^{2mp})\,ds\\ &\leqq t^{2m-1}c^{2m}2^{2m-1}\left[1+\sup_{0\leqq s\leqq t}|x^\varepsilon(s)|^{2mp}\right]\\ &\quad\times\exp\left[\frac{-2m\kappa t}{\varepsilon}\right]\int_0^t\exp\left[\frac{2m\kappa s}{\varepsilon}\right]ds\\ &= t^{2m-1}c^{2m}2^{2m-1}\left[1+\sup_{0\leqq s\leqq t}|x^\varepsilon(s)|^{2mp}\right]\\ &\quad\times\frac{\varepsilon}{2m\kappa}\left[1-\exp\left[\frac{-2m\kappa t}{\varepsilon}\right]\right].\end{aligned}$$

By the moment condition on ϕ we notice that

$$E[V(\phi)^{2mp}]\leqq(E[V(\phi)^{mq}])^{2p/q}<\infty \quad \text{for } m\geqq 2.$$

Thus the estimate (3.6) of Lemma 3.4 holds with m replaced by mp. So

$$\begin{aligned}E\left[\sup_{0\leqq t\leqq T}|I_2^\varepsilon(t)|^{2m}\right]&\leqq c^{2m}2^{2m-1}T^{2m-1}[1+K_{mp}(T)]\\ &\quad\times\frac{\varepsilon}{2m\kappa}\left[1-\exp\left[\frac{-2m\kappa T}{\varepsilon}\right]\right],\end{aligned}\tag{4.5}$$

where $K_{mp}(T)$ is the constant obtained by $K_m(T)$ in (3.6) with m replaced by mp.

Next we evaluate $|I_3^\varepsilon(t)|^{2m}$. Use the inequality that

$$E\left[\left(\int_0^t h(u)\,dw(u)\right)^{2m}\right]\leqq[m(2m-1)]^m t^{m-1}\int_0^t E[h(u)^{2m}]\,du$$

for a nonanticipating Brownian functional $h(u)$. Then

$$E[|I_3^\varepsilon(t)|^{2m}]\leqq J_m^\varepsilon(t)\quad\text{for all } t\geqq 0,\tag{4.6}$$

where

$$J_m^\varepsilon(t)=\delta^{2m}[m(2m-1)]^m t^{m-1}\frac{\varepsilon}{2m\kappa}\left[1-\exp\left[\frac{-2m\kappa t}{\varepsilon}\right]\right].$$

Notice that $I_3^\varepsilon(t)$ is the pathwise unique solution of the following Langevin equation:

$$dI(t)=-\frac{\kappa}{\varepsilon}I(t)\,dt+\delta\,dw(t),\qquad I(0)=0.$$

Put

$$M(t)=|I(t)|^{2m}+\frac{2m\kappa}{\varepsilon}\int_0^t|I(s)|^{2m}\,ds.$$

Then the Ito formula applies to the stochastic differential $d[|I(t)|^{2m}]$ with the following result:

$$M(t)=\int_0^t \delta^2 m(2m-1)|I(s)|^{2m-2}\,ds+2m\delta\int_0^t |I(s)|^{2m-1}\,dw(s).$$

Namely, $M(t)$ is a nonnegative submartingale, which yields

$$E\left[\sup_{0\leqq t\leqq T}|I(t)|^{2m}\right]\leqq E\left[\sup_{0\leqq t\leqq T}M(t)\right]\leqq 2(E[M(T)^2])^{1/2}.$$

Here, by the definition of $M(t)$ we see

$$E[M(T)^2]\leqq 2\left[\delta^4[m(2m-1)]^2T\int_0^T E[|I(s)|^{4m-4}]\,ds\right.$$
$$\left.+\delta^2[2m]^2\int_0^T E[|I(s)|^{4m-2}]\,ds\right].$$

Let m be such that $m\geqq 2$. Then, by the definition of $J^{\varepsilon}_m(t)$ in (4.6), both $J^{\varepsilon}_{2m-2}(t)$ and $J^{\varepsilon}_{2m-1}(t)$ are well defined. Further,

$$J^{\varepsilon}_{2m-2}(t)\leqq J^{\varepsilon}_{2m-2}(T)\quad\text{for all } t\leqq T,$$
$$J^{\varepsilon}_{2m-1}(t)\leqq J^{\varepsilon}_{2m-1}(T)\quad\text{for all } t\leqq T.$$

So the estimate (4.6) yields

$$E\left[\sup_{0\leqq t\leqq T}|I(t)|^{2m}\right]\leqq 2\sqrt{2}[\delta^2 m(2m-1)T(J^{\varepsilon}_{2m-2}(T))^{1/2}\tag{4.7}$$
$$+\delta 2m(TJ^{\varepsilon}_{2m-1}(T))^{1/2}],$$

since $(a+b)^{1/2}\leqq a^{1/2}+b^{1/2}$ for $a\geqq 0$ and $b\geqq 0$. Observe that for $n\geqq 1$

$$J^{\varepsilon}_n(t)^{1/2}\leqq \delta^n[n(2n-1)]^{n/2}T^{(n-1)/2}\left(\frac{\varepsilon}{2n\kappa}\right)^{1/2}.$$

Take expectation on (4.4). Then, by (4.5) and (4.7), we get (4.3) as follows:

$$E\left[\sup_{0\leqq t\leqq T}|Y^{\varepsilon}(t)|^{2m}\right]$$
$$\leqq 2^{2m-1}\left[c^{2m}2^{2m-1}T^{2m-1}[1+K_{mp}(T)]\left(\frac{1}{2m\kappa}\right)\varepsilon+H_m\delta^{2m}T^{m-1/2}\left(\frac{1}{2m\kappa}\right)^{1/2}\sqrt{\varepsilon}\right],$$

where

$$H_m=2\sqrt{2}[m(2m-1)[(2m-2)(4m-5)]^{m-1}+2m[(2m-1)(4m-3)]^{m-1/2}].$$

Now we show (4.3)′. Since $Y^{\varepsilon}(t)=-I^{\varepsilon}_2(t)+I^{\varepsilon}_3(t)$, we see that for $t\geqq s$,

$$Y^{\varepsilon}(t)-Y^{\varepsilon}(s)=-(I^{\varepsilon}_2(t)-I^{\varepsilon}_2(s))+(I^{\varepsilon}_3(t)-I^{\varepsilon}_3(s)),$$

and so

$$E[|Y^{\varepsilon}(t)-Y^{\varepsilon}(s)|^4]\leqq 2^3[E[|I^{\varepsilon}_2(t)-I^{\varepsilon}_2(s)|^4]+E[|I^{\varepsilon}_3(t)-I^{\varepsilon}_3(s)|^4]].$$

Here the estimate (6.2) of Lemma 6.1 in § 6 is applied to $I^{\varepsilon}_3(t)$:

$$\sup_{0<\varepsilon\leqq 1}E[|I^{\varepsilon}_3(t)-I^{\varepsilon}_3(s)|^4]\leqq 15\delta^4(t-s)^2\quad\text{for } 0\leqq s\leqq t.$$

Thus, in order to prove (4.3)′ we have only to show

$$\sup_{0<\varepsilon\leqq 1} E[|I_2^\varepsilon(t)-I_2^\varepsilon(s)|^4]\leqq K(t-s)^2 \quad \text{for } 0\leqq s\leqq t\leqq T$$

with a constant $K>0$ independent of ε.

By the assumption, since $|g(x)|\leqq c(1+|x|^p)$ for $x\in R^1$, we see that for $0\leqq u\leqq T$,

$$\begin{aligned}|I_2^\varepsilon(u)| &= \left|\exp\left[\frac{-\kappa u}{\varepsilon}\right]\int_0^u \exp\left[\frac{\kappa v}{\varepsilon}\right] g(x^\varepsilon(v))\,dv\right| \\ &\leqq c\left(1+\sup_{0\leqq v\leqq T}|x^\varepsilon(v)|^p\right)\exp\left[\frac{-\kappa u}{\varepsilon}\right]\int_0^u \exp\left[\frac{\kappa v}{\varepsilon}\right]dv \\ &\leqq \varepsilon\left(\frac{c}{\kappa}\right)\left(1+\sup_{0\leqq v\leqq T}|x^\varepsilon(v)|^p\right).\end{aligned}$$

On the other hand, $I_2^\varepsilon(t)$ satisfies the linear ordinary differential equation of first order:

$$\frac{d}{dt}I_2^\varepsilon(t)=-\frac{\kappa}{\varepsilon}I_2^\varepsilon(t)+g(x^\varepsilon(t)) \quad \text{with } I_2^\varepsilon(0)=0.$$

So, for $0\leqq s\leqq t\leqq T$,

$$I_2^\varepsilon(t)-I_2^\varepsilon(s)=-\frac{\kappa}{\varepsilon}\int_s^t I_2^\varepsilon(u)\,du+\int_s^t g(x^\varepsilon(u))\,du$$

and

$$\begin{aligned}|I_2^\varepsilon(t)-I_2^\varepsilon(s)| &\leqq \frac{\kappa}{\varepsilon}\int_s^t |I_2^\varepsilon(u)|\,du+\int_s^t |g(x^\varepsilon(u))|\,du \\ &\leqq \frac{\kappa}{\varepsilon}\times\varepsilon\left(\frac{c}{\kappa}\right)\left(1+\sup_{0\leqq v\leqq T}|x^\varepsilon(v)|^p\right)(t-s) \\ &\quad +c\left(1+\sup_{0\leqq v\leqq T}|x^\varepsilon(v)|^p\right)(t-s) \\ &=2c\left(1+\sup_{0\leqq v\leqq T}|x^\varepsilon(v)|^p\right)(t-s).\end{aligned}$$

Under the assumption, the estimate (3.6) holds with $m=2p$, that is,

$$E\left[\sup_{0\leqq v\leqq T}|x^\varepsilon(v)|^{4p}\right]\leqq K_{2p}(T).$$

Therefore,

$$\begin{aligned}E[|I_2^\varepsilon(t)-I_2^\varepsilon(s)|^4] &\leqq 2^4c^4E\left[\left(1+\sup_{0\leqq v\leqq T}|x^\varepsilon(v)|^p\right)^4\right](t-s)^4 \\ &\leqq K(t-s)^2 \quad \text{for } 0\leqq s\leqq t\leqq T\end{aligned}$$

with a constant $K>0$ independent of ε. This inequality together with Lemma 6.1 imply (4.3)′. Hence the proof is complete.

LEMMA 4.3. *Under the same assumptions as in Lemma* 4.2, *let* $X^\varepsilon(t)$ *be the process defined by* (2.1). *Let* $T<\infty$ *be arbitrary and fixed. Then*

$$\sup_{0<\varepsilon\leqq 1} E[|X^\varepsilon(t)-X^\varepsilon(s)|^4]\leqq D'(t-s)^2 \tag{4.8}$$

for $0\leqq s\leqq t\leqq T$ *with a constant* $D'>0$ *independent of* ε.

Proof. By (2.1)′, $(X^\varepsilon(t),\ Y^\varepsilon(t))$ satisfies

$$\kappa(X^\varepsilon(t)-X^\varepsilon(s))=(y^\varepsilon(t)-y^\varepsilon(s))-(Y^\varepsilon(t)-Y^\varepsilon(s))$$

for $0\leqq s\leqq t\leqq T$, and hence

$$\kappa^4|X^\varepsilon(t)-X^\varepsilon(s)|^4\leqq 2^3[|y^\varepsilon(t)-y^\varepsilon(s)|^4+|Y^\varepsilon(t)-Y^\varepsilon(s)|^4]$$

for $0\leqq s\leqq t\leqq T$. Take expectation on both sides of the above equation and apply the estimates (4.1) of Lemma 4.1 and (4.3)′ of Lemma 4.2. Then we obtain (4.8). Hence the proof is complete.

Remark 4.1. Lemmas 4.1 and 4.3, as follows from Prokhorov [6], assure the relative compactness of the families of probability measures induced by $y^\varepsilon=\{y^\varepsilon(t)\}$ and $X^\varepsilon=\{X^\varepsilon(t)\}$. By the representation theorem of Skorokhod [7], without loss of generality, we can assume that there exist a subsequence $\{\varepsilon_j\}_{j=1,2,\dots}$ of $\{\varepsilon\}$ and random processes $\tilde{y}(t)$ and $\tilde{X}(t)$, such that

$$y^{\varepsilon_j}(t)\to\tilde{y}(t)\quad\text{and}\quad X^{\varepsilon_j}(t)\to\tilde{X}(t)\quad\text{with probability 1}$$

uniformly for each finite time interval as $j\to\infty$.

LEMMA 4.4. *Under the same assumptions as in Lemma* 4.2, *let* $\{\varepsilon_j\}_{j=1,2,\dots}$ *and* $\tilde{X}(t)$ *be a subsequence and a process for which Remark* 4.1 *holds. For the sake of simplicity, let us consider that* $\{\varepsilon_j\}_{j=1,2,\dots}=\{\varepsilon\}$. *Then, for each* $T<\infty$ *there exists a constant* $\tilde{C}>0$ *independent of* ε, *satisfying the following estimates*:

$$E\left[\sup_{0\leqq t\leqq T}|X^\varepsilon(t)|^{mq}\right]\leqq\tilde{C},\qquad q=\max\{2p,2\}, \tag{4.9}$$

$$E\left[\sup_{0\leqq t\leqq T}|g(X^\varepsilon(t))|^{2m}\right]\leqq\tilde{C}, \tag*{(4.9)′}$$

$$E\left[\sup_{0\leqq t\leqq T}|g(\tilde{X}(t))|^{2m}\right]\leqq\tilde{C}, \tag*{(4.9)″}$$

where $m\geqq 2$ *and* $p\geqq 1$ *are as in the hypothesis of Lemma* 4.2.

Proof. We shall denote various positive constants independent of ε by the same symbol $\tilde{C}$. By the assumption, since $|g(x)|\leqq c(1+|x|^p)$ for all $x\in R^1$, $g(x)$ satisfies

$$|g(x)|^{2m}\leqq\tilde{C}(1+|x|^{mq})\quad\text{for all } x\in R^1\quad\text{and}\quad m\geqq 1 \tag{4.10}$$

with $q=\max\{2p,2\}$. Recall (2.1), so that

$$\begin{aligned}|X^\varepsilon(t)|^{mq}&=\left|x^\varepsilon(t)-\left(\xi-\frac{\eta}{\kappa}\right)\exp\left[\frac{-\kappa t}{\varepsilon}\right]\right|^{mq}\\&\leqq\tilde{C}[|x^\varepsilon(t)|^{mq}+|\phi|^{mq}],\end{aligned}$$

where $\phi=(\xi,\eta)$ is the initial vector in the hypothesis. Thus

$$\sup_{0\leqq t\leqq T}|X^\varepsilon(t)|^{mq}\leqq\tilde{C}\left[\sup_{0\leqq t\leqq T}|x^\varepsilon(t)|^{mq}+|\phi|^{mq}\right]. \tag{4.11}$$

The condition (iii) of Assumption 2.2 and the moment condition on ϕ imply

$$E[|\phi|^{mq}] \leqq (E[|\phi|^{2mq}])^{1/2} \leqq l^{mq/2}(E[V(\phi)^{mq}])^{1/2} < \infty, \tag{4.12}$$

where $m \geqq 2$. Therefore, the estimate (3.6) of Lemma 3.4 holds with m replaced by $mq/2$. Namely,

$$E\left[\sup_{0 \leqq t \leqq T} |x^{\varepsilon}(t)|^{mq}\right] \leqq \tilde{C}. \tag{4.13}$$

Take expectation on (4.11). Then, by (4.12) and (4.13) we get (4.9). So, by (4.9) and (4.10) we get (4.9)'. Consider that

$$g(X^{\varepsilon}(t)) \to g(\tilde{X}(t)) \quad \text{with probability 1} \quad \text{as } \varepsilon \to 0.$$

Then, by (4.9)' and the Fatou lemma we obtain (4.9)''. Hence the proof is complete.

5. Proof of Theorem 2.1. First of all we consider the one-dimensional SDE (2.2).

LEMMA 5.1. *Suppose that $g(x)$ is a once continuously differentiable function satisfying*

$$-xg(x) \leqq \alpha \quad \textit{for all } x \in R^1 \quad \textit{with a constant } \alpha > 0.$$

Let $\phi = (\xi, \eta)$ be a two-dimensional random vector independent of the two-dimensional Brownian motion process, such that

$$E[|\phi|^2] < \infty.$$

Then there exists a pathwise unique solution $X(t)$ of (2.2) with the initial state $X(0) = \eta/\kappa$. Moreover, suppose that $g(x)$ satisfies all conditions of Assumption 2.2 and that

$$E[V(\phi)^{2m}] < \infty \quad \textit{for an integer } m \geqq 1.$$

Then

$$E[(1+\tfrac{1}{2}|X(t)|^2)^{2m}] \leqq E[(1+\tfrac{1}{2}|X(0)|^2)^{2m}] \exp[b_m t], \tag{5.1}$$

where

$$b_m = 2m\left[\frac{\alpha}{\kappa} + \left(2m - \frac{1}{2}\right)\left(\frac{\delta}{\kappa}\right)^2\right].$$

Proof. Denote by L the differential generator associated with SDE (2.2):

$$L = -\frac{1}{\kappa} g(x) \frac{d}{dx} + \frac{1}{2}\left(\frac{\delta}{\kappa}\right)^2 \frac{d^2}{dx^2}, \qquad x \in R^1.$$

Then, the assumption on $g(x)$ implies

$$L\left[\frac{1}{2}x^2\right] = -\frac{1}{\kappa} xg(x) + \frac{1}{2}\left(\frac{\delta}{\kappa}\right)^2 \leqq \frac{\alpha}{\kappa} + \frac{1}{2}\left(\frac{\delta}{\kappa}\right)^2$$

for all $x \in R^1$. Moreover,

$$\frac{1}{2}x^2 \to \infty \quad \text{as } |x| \to \infty.$$

So, any solution $X(t)$ of (2.2) with the initial state $X(0) = \eta/\kappa$ such that $E[|\eta|^2] < \infty$, as follows from Hasminskii [2] and Narita [4], cannot explode. Hence the pathwise uniqueness holds for the solution of (2.2).

Next it is easy to see that

$$L[(1+\tfrac{1}{2}x^2)^{2m}] \leqq b_m(1+\tfrac{1}{2}x^2)^{2m} \quad \text{for all } x \in R^1$$

with the constant

$$b_m = 2m\left[\frac{\alpha}{\kappa} + \left(2m - \frac{1}{2}\right)\left(\frac{\delta}{\kappa}\right)^2\right].$$

Recall the condition (iii) of Assumption 2.2 and the moment condition on ϕ, so that

$$E\left[\left(\frac{1}{2}|X(0)|^2\right)^{2m}\right] \leqq E\left[\left(\frac{|\phi|^2}{2\kappa^2}\right)^{2m}\right] \leqq \left(\frac{l}{2\kappa^2}\right)^{2m} E[V(\phi)^{2m}] < \infty.$$

Then, evaluating $E[(1+\frac{1}{2}|X(t)|^2)^{2m}]$ by the Ito formula and the Gronwall inequality we obtain (5.1). Hence the proof is complete.

Proof of Theorem 2.1. Let m be the integer for which Assumption 2.3 holds:

$$E[V(\phi)^{mq}] < \infty, \qquad q = \max\{2p, 2\}, \quad m \geqq 2.$$

Then the initial vector ϕ satisfies all moment conditions in the hypotheses of Lemmas 3.1, 3.4, 4.1, 4.4, and 5.1 automatically. Under Assumptions 2.1 and 2.2, the function $g(x)$ satisfies all growth conditions in the hypotheses of Lemmas 3.1, 3.4, 4.1, 4.4, and 5.1. Therefore, the family $\{y^\varepsilon(t), x^\varepsilon(t), Y^\varepsilon(t), X^\varepsilon(t), X(t)\}$ of processes satisfies the family $\{(3.5), (3.6), (4.3), (4.9), (5.1)\}$ of estimates, respectively.

Denote by $\{\varepsilon_j\}_{j=1,2,\ldots}$ a subsequence of $\{\varepsilon\}$ for which Remark 4.1 holds. For simplicity of the notation, let us consider that $\{\varepsilon_j\} = \{\varepsilon\}$. Let $\tilde{y}(t)$ and $\tilde{X}(t)$ be the random processes for which Remark 4.1 holds. Then we proceed to the proof by showing the following results.

Step 1. $\tilde{y}(t) = \kappa\tilde{X}(t)$ for all $t \geqq 0$ with probability 1.

Step 2. There exists a one-dimensional Brownian motion process $\tilde{w}(t)$ such that

$$\tilde{y}(t) = \eta - \int_0^t g(\tilde{X}(s))\,ds + \delta\tilde{w}(t) \quad \text{for all } t \geqq 0$$

with probability 1.

Step 3. $$\tilde{X}(t) = \frac{\eta}{\kappa} - \frac{1}{\kappa}\int_0^t g(\tilde{X}(s))\,ds + \frac{\delta}{\kappa}\tilde{w}(t) \quad \text{for all } t \geqq 0$$

with probability 1.

Step 4. $X^\varepsilon(t)$ converges weakly in $C([0, T]; R^1)$, $T < \infty$, but arbitrary, as $\varepsilon \to 0$ to the solution $X(t)$ governed by SDE (2.2) with the initial state $X(0) = \eta/\kappa$.

Proof of Step 1. Observe that

$$\begin{aligned} E[|\tilde{y}(t) - \kappa\tilde{X}(t)|] \leqq{} & E[|\tilde{y}(t) - y^\varepsilon(t)|] + E[|Y^\varepsilon(t)|] \\ & + \kappa E[|X^\varepsilon(t) - \tilde{X}(t)|], \end{aligned} \tag{5.2}$$

where $Y^\varepsilon(t) = y^\varepsilon(t) - \kappa X^\varepsilon(t)$. Since the estimates (3.5) of Lemma 3.3 and (4.9) of Lemma 4.4 hold for $y^\varepsilon(t)$ and $X^\varepsilon(t)$ with $m \geqq 2$, both $\{y^\varepsilon(t)\}$ and $\{X^\varepsilon(t)\}$ are equi-integrable with respect to ε, while $y^\varepsilon(t) \to \tilde{y}(t)$ and $X^\varepsilon(t) \to \tilde{X}(t)$ with probability 1 as $\varepsilon \to 0$. Thus the L^1-convergence of $y^\varepsilon(t)$ and $X^\varepsilon(t)$ to $\tilde{y}(t)$ and $\tilde{X}(t)$ holds, respectively. On the other hand, the estimate (4.3) of Lemma 4.2 implies the L^{2m}-convergence of $Y^\varepsilon(t)$ to zero as $\varepsilon \to 0$. Thus, letting $\varepsilon \to 0$ in (5.2), we have

$$E[|\tilde{y}(t) - \kappa\tilde{X}(t)|] = 0,$$

and hence

$$P(\tilde{y}(t)=\kappa\tilde{X}(t))=1 \quad \text{for each } t\geqq 0.$$

Since $\tilde{y}(t)$ and $\tilde{X}(t)$ are continuous with probability 1, the result of Step 1 holds.

Proof of Step 2. Put

$$\Delta(t)=\tilde{y}(t)-\left[\eta-\int_0^t g(\tilde{X}(s))\,ds\right].$$

Rewrite $\Delta(t)$ as

$$\Delta(t)=[\tilde{y}(t)-y^\varepsilon(t)]+\left[y^\varepsilon(t)-\left(\eta-\int_0^t g(\tilde{X}(s))\,ds\right)\right].$$

Then, since $y^\varepsilon(t)=\eta-\int_0^t g(x^\varepsilon(s))\,ds+\delta w(t)$, we see

$$\Delta(t)=\Delta_1^\varepsilon(t)-\Delta_2^\varepsilon(t)+\delta w(t), \tag{5.3}$$

where

$$\Delta_1^\varepsilon(t)=\tilde{y}(t)-y^\varepsilon(t) \quad\text{and}\quad \Delta_2^\varepsilon(t)=\int_0^t [g(x^\varepsilon(s))-g(\tilde{X}(s))]\,ds.$$

Since the estimate (3.5) of Lemma 3.3 holds with $m\geqq 2$, $\{|y^\varepsilon(t)|^2\}$ is equi-integrable with respect to ε, while $y^\varepsilon(t)\to\tilde{y}(t)$ with probability 1 as $\varepsilon\to 0$. Thus

$$E[|\Delta_1^\varepsilon(t)|^2]\to 0 \quad\text{as } \varepsilon\to 0. \tag{5.4}$$

On the other hand, the Schwarz inequality yields

$$|\Delta_2^\varepsilon(t)|^2\leqq t\int_0^t |g(x^\varepsilon(s))-g(\tilde{X}(s))|^2\,ds.$$

For a moment suppose that the following result holds:

$$E\left[\int_0^t |g(x^\varepsilon(s))-g(\tilde{X}(s))|^2\,ds\right]\to 0 \quad\text{as } \varepsilon\to 0. \tag{5.5}$$

Then we get

$$E[|\Delta_2^\varepsilon(t)|^2]\to 0 \quad\text{as } \varepsilon\to 0. \tag{5.6}$$

So, taking expectation on (5.3), by (5.4) and (5.6) we have

$$|E[\Delta(t)]|\leqq|E[\Delta_1^\varepsilon(t)]|+|E[\Delta_2^\varepsilon(t)]|\to 0 \quad\text{as } \varepsilon\to 0,$$

and hence

$$E[\Delta(t)]=0.$$

Further, it follows from (5.3) that

$$\Delta(t)^2=(\Delta_1^\varepsilon(t)-\Delta_2^\varepsilon(t))^2+2\delta(\Delta_1^\varepsilon(t)-\Delta_2^\varepsilon(t))w(t)+\delta^2 w(t)^2. \tag{5.7}$$

Here (5.4) and (5.6) imply

$$E[(\Delta_1^\varepsilon(t)-\Delta_2^\varepsilon(t))^2]\leqq 2[E[|\Delta_1^\varepsilon(t)|^2]+E[|\Delta_2^\varepsilon(t)|^2]]$$

$$\to 0 \quad \text{as } \varepsilon\to 0,$$

$$|E[(\Delta_1^\varepsilon(t)-\Delta_2^\varepsilon(t))w(t)]|\leqq (E[(\Delta_1^\varepsilon(t)-\Delta_2^\varepsilon(t))^2])^{1/2}t^{1/2}$$

$$\to 0 \quad \text{as } \varepsilon\to 0.$$

Hence, taking expectation on (5.7) we get

$$\begin{aligned}E[\Delta(t)^2]&=E[(\Delta_1^\varepsilon(t)-\Delta_2^\varepsilon(t))^2]\\&\quad+2\delta E[(\Delta_1^\varepsilon(t)-\Delta_2^\varepsilon(t))w(t)]+\delta^2 t\\&\to \delta^2 t \quad \text{as } \varepsilon\to 0.\end{aligned}$$

Namely,

$$E[\Delta(t)^2]=\delta^2 t.$$

Since $\tilde{y}(t)$ and $\tilde{X}(t)$ are continuous, $\Delta(t)$ is also continuous. Define $\tilde{w}(t)$ by

$$\tilde{w}(t)=\frac{1}{\delta}\Delta(t).$$

Then, according to the above argument, $\{\tilde{w}(t)\}$ is a continuous square integrable martingale with $E[\tilde{w}(t)]=0$ and $E[\tilde{w}(t)^2]=t$. Thus, $\tilde{w}(t)$ is a one-dimensional Brownian motion process. Hence the result of Step 2 is valid.

Therefore, in order to prove Step 2 we have only to show (5.5). Since $g(x)$ is differentiable, the mean value theorem yields

$$g(X+h)=g(X)+g'(\zeta)h, \qquad \zeta=X+\theta h, \quad 0<\theta<1.$$

Consider the expression (2.1)′, so that

$$x^\varepsilon(s)=X^\varepsilon(s)+h^\varepsilon(s) \quad \text{with } h^\varepsilon(s)=\left(\xi-\frac{\eta}{\kappa}\right)\exp\left[\frac{-\kappa s}{\varepsilon}\right].$$

Then

$$g(x^\varepsilon(s))=g(X^\varepsilon(s))+g'(\zeta^\varepsilon(s))h^\varepsilon(s),$$

where

$$\zeta^\varepsilon(s)=X^\varepsilon(s)+\theta^\varepsilon(s)h^\varepsilon(s), \qquad 0<\theta^\varepsilon(s)<1.$$

Observe that

$$\begin{aligned}g(x^\varepsilon(s))-g(\tilde{X}(s))&=[g(x^\varepsilon(s))-g(X^\varepsilon(s))]+[g(X^\varepsilon(s))-g(\tilde{X}(s))]\\&=g'(\zeta^\varepsilon(s))h^\varepsilon(s)+[g(X^\varepsilon(s))-g(\tilde{X}(s))],\end{aligned}$$

and so

$$(5.8)\quad |g(x^\varepsilon(s))-g(\tilde{X}(s))|^2\leqq 2[|g'(\zeta^\varepsilon(s))|^2|h^\varepsilon(s)|^2+|g(X^\varepsilon(s))-g(\tilde{X}(s))|^2].$$

Then we first evaluate the right-hand side of (5.8). By Assumption 2.1, since $|g'(\zeta)|\leqq c(1+|\zeta|^p)$ for all $\zeta\in R^1$, we see

$$|g'(\zeta^\varepsilon(s))||h^\varepsilon(s)|\leqq c[1+2^{p-1}(|X^\varepsilon(s)|^p+|h^\varepsilon(s)|^p)].$$

In the following we denote various positive constants independent of ε by the same symbol A. Notice the inequalities such that

$$|h^\varepsilon(s)|^2 \leqq A|\phi|^2 \exp\left[\frac{-2\kappa s}{\varepsilon}\right],$$

$$|h^\varepsilon(s)|^{2p} \leqq A|\phi|^{2p} \exp\left[\frac{-2\kappa p s}{\varepsilon}\right],$$

$$|g'(\zeta^\varepsilon(s))|^2|h^\varepsilon(s)|^2 \leqq A\left[|\phi|^2+|X^\varepsilon(s)|^{2p}|\phi|^2 + |\phi|^{2p+2}\exp\left[\frac{-2\kappa p s}{\varepsilon}\right]\right]\exp\left[\frac{-2\kappa s}{\varepsilon}\right].$$

Use the inequality such that $xy \leqq (x^2+y^2)/2$. Then

$$\begin{aligned}(5.9)\qquad &|g'(\zeta^\varepsilon(s))|^2|h^\varepsilon(s)|^2\\ &\leqq A\left[|\phi|^2+|X^\varepsilon(s)|^{4p}+|\phi|^4+|\phi|^{2p+2}\exp\left[\frac{-2\kappa p s}{\varepsilon}\right]\right]\\ &\times\exp\left[\frac{-2\kappa s}{\varepsilon}\right].\end{aligned}$$

By condition (iii) of Assumption 2.2, ϕ satisfies

$$|\phi|^2 \leqq lV(\phi), \qquad |\phi|^4 \leqq l^2V(\phi)^2,$$

$$\begin{aligned}|\phi|^{2p+2} = |\phi|^{2p}|\phi|^2 &\leqq \tfrac{1}{2}[|\phi|^{4p}+|\phi|^4]\\ &\leqq \tfrac{1}{2}[l^{2p}V(\phi)^{2p}+l^2V(\phi)^2].\end{aligned}$$

Recall Assumption 2.3: $E[V(\phi)^{mq}]<\infty$, $q=\max\{2p,2\}$, $m\geqq 2$. Then, since $2\leqq 2p\leqq q<mq$ for $m\geqq 2$, the Hölder inequality yields

$$E[V(\phi)^2] \leqq (E[V(\phi)^{2p}])^{1/p} \leqq (E[V(\phi)^{mq}])^{2/(mq)}<\infty,$$

and hence

$$E[|\phi|^2] \leqq (E[|\phi|^4])^{1/2} \leqq (E[|\phi|^{2p+2}])^{2/(p+1)}<\infty.$$

Further, the estimate (4.9) of Lemma 4.4 holds for $X^\varepsilon(s)$ with the following result:

$$E\left[\sup_{0\leqq s\leqq t}|X^\varepsilon(s)|^{4p}\right] \leqq 1+E\left[\sup_{0\leqq s\leqq t}|X^\varepsilon(s)|^{mq}\right]<\infty,$$

where $4p\leqq 2q\leqq mq$ for $m\geqq 2$. Therefore, taking expectation on (5.9) we have

$$(5.10)\qquad E[|g'(\zeta^\varepsilon(s))|^2|h^\varepsilon(s)|^2] \leqq A\exp\left[\frac{-2\kappa s}{\varepsilon}\right] \quad \text{for all } 0\leqq s\leqq t.$$

Next, set

$$\lambda^\varepsilon(t) = E\left[\int_0^t |g(X^\varepsilon(s))-g(\tilde{X}(s))|^2\,ds\right].$$

Notice that the estimates (4.9)′ and (4.9)″ of Lemma 4.4 hold for $g(X^\varepsilon(s))$ and $g(\tilde{X}(s))$. Then, since $g(X^\varepsilon(s)) \to g(\tilde{X}(s))$ with probability 1 as $\varepsilon \to 0$, by the dominated convergence theorem we get

$$\lambda^\varepsilon(t) \to 0 \quad \text{as } \varepsilon \to 0. \tag{5.11}$$

Accordingly, by (5.8), (5.10), and (5.11) we obtain

$$\begin{aligned} &E\left[\int_0^t |g(x^\varepsilon(s)) - g(\tilde{X}(s))|^2\, ds\right] \\ &\leqq 2\left[A\int_0^t \exp\left[\frac{-2\kappa s}{\varepsilon}\right] ds + \lambda^\varepsilon(t)\right] \\ &= 2\left[A\frac{\varepsilon}{2\kappa}\left(1 - \exp\left[\frac{-2\kappa t}{\varepsilon}\right]\right) + \lambda^\varepsilon(t)\right] \to 0 \quad \text{as } \varepsilon \to 0. \end{aligned}$$

Thus (5.5) holds. Hence the proof of Step 2 is complete.

Proof of Step 3. Obviously, the result of Step 3 follows from Steps 1 and 2.

Proof of Step 4. By virtue of Step 3, the limit process $\tilde{X}(t)$ has the same law with the solution $X(t)$ of SDE (2.2) with the initial state $X(0) = \eta/\kappa$. According to Lemma 5.1, the pathwise uniqueness holds for the solution of (2.2), which implies the law uniqueness. So the result of Step 4 holds. Moreover, by (4.3) of Lemma 4.2, $Y^\varepsilon(t)$ satisfies

$$E\left[\sup_{0 \leqq t \leqq T} |Y^\varepsilon(t)|^{2m}\right] \to 0 \quad \text{as } \varepsilon \to 0.$$

Steps 1–4 also imply the weak convergence of $y^\varepsilon(t)$ to $\kappa X(t)$ as $\varepsilon \to 0$. Hence the proof of Theorem 2.1 is complete.

Remark 5.1. In the proof of (5.5), a process $\theta^\varepsilon(t)$ arises from the mean value theorem. Let $\{\varepsilon_j\}_{j=1,2,\ldots}$ be a subsequence for which Remark 4.1 holds. For notational simplicity, and without loss of generality, we will usually *not* use the tilde notation, and simply assume that

$$X^{\varepsilon_j}(t) \to X(t) \text{ with probability 1 as } j \to \infty,$$

where $X(t)$ satisfies (2.2) and consider that $\{\varepsilon_j\} = \{\varepsilon\}$. Suppose that $g(x)$ is a twice continuously differentiable function satisfying

$$g''(x) \neq 0 \quad \text{for all } x \in R^1.$$

Then, for each $t \neq 0$,

$$\theta^\varepsilon(t) \to \tfrac{1}{2} \quad \text{on } \{\eta \neq \kappa\xi\} \quad \text{with probability 1} \quad \text{as } \varepsilon \to 0, \tag{5.12}$$

where $\phi = (\xi, \eta)$ is the initial vector given in (1.2).

We will show (5.12). In fact, by the theorem of Taylor, the following equations hold:

$$\begin{aligned} g(a+h) &= g(a) + hg'(a+\theta h) \\ &= g(a) + h\{g'(a) + \theta h g''(a + \theta_1\theta h)\} \\ &= g(a) + hg'(a) + \theta h^2 g''(a + \theta_1 \theta h), \end{aligned}$$

where $0 < \theta < 1$ and $0 < \theta_1 < 1$, and

$$g(a+h) = g(a) + hg'(a) + \tfrac{1}{2}h^2 g''(a + \theta_2 h),$$

where $0<\theta_2<1$. Thus

$$\theta h^2 g''(a+\theta_1\theta h)=\tfrac{1}{2}h^2 g''(a+\theta_2 h).$$

Consider that

$$a=X^\varepsilon(t),\qquad h=h^\varepsilon(t)=\left(\xi-\frac{\eta}{\kappa}\right)\exp\left[\frac{-\kappa t}{\varepsilon}\right],$$

$$\theta=\theta^\varepsilon(t)\quad\text{and}\quad\theta_i=\theta_i^\varepsilon(t),\quad\text{where } i=1,2.$$

Then, since $\exp[-\kappa t/\varepsilon]\neq 0$ for all $t\geqq 0$, we have

$$\theta^\varepsilon(t)\left(\xi-\frac{\eta}{\kappa}\right)^2 g''(X^\varepsilon(t)+\theta_1^\varepsilon(t)\theta^\varepsilon(t)h^\varepsilon(t))$$

$$=\frac{1}{2}\left(\xi-\frac{\eta}{\kappa}\right)^2 g''(X^\varepsilon(t)+\theta_2^\varepsilon(t)h^\varepsilon(t)),\quad\text{where } t\geqq 0.$$

Note that for each $t\geqq 0$,

$$X^\varepsilon(t)\to X(t)\quad\text{with probability 1}\quad\text{as } \varepsilon\to 0.$$

Observe that for each $t\neq 0$,

$$\exp\left[\frac{-\kappa t}{\varepsilon}\right]\to 0\quad\text{as } \varepsilon\to 0,$$

which implies that for each $t\neq 0$,

$$\theta_1^\varepsilon(t)\theta^\varepsilon(t)h^\varepsilon(t)\to 0\quad\text{and}\quad\theta_2^\varepsilon(t)h^\varepsilon(t)\to 0$$

with probability 1 as $\varepsilon\to 0$. Then, since $g''(x)$ is continuous in x, we see that for $t\neq 0$,

$$g''(X^\varepsilon(t)+\theta_1^\varepsilon(t)\theta^\varepsilon(t)h^\varepsilon(t))\to g''(X(t))\quad\text{with probability 1}\quad\text{as } \varepsilon\to 0$$

and

$$g''(X^\varepsilon(t)+\theta_2^\varepsilon(t)h^\varepsilon(t))\to g''(X(t))\quad\text{with probability 1}\quad\text{as } \varepsilon\to 0.$$

Therefore, for each $t\neq 0$,

$$\left(\lim_{\varepsilon\to 0}\theta^\varepsilon(t)\right)\left(\xi-\frac{\eta}{\kappa}\right)^2 g''(X(t))=\frac{1}{2}\left(\xi-\frac{\eta}{\kappa}\right)^2 g''(X(t))\quad\text{on } \{\eta\neq\kappa\xi\},$$

from which follows (5.12).

6. Appendix (a priori bounds of the solution of the Langevin equation). In order to prove the estimate (4.3)′ of Lemma 4.2 we shall need Lemma 6.1. Let us consider the solution $I(t)$ of the one-dimensional Langevin equation with a drift multiplied by $1/\varepsilon$:

$$dI(t)=-\frac{\kappa}{\varepsilon}I(t)\,dt+\delta\,dw(t)\quad\text{with } I(0)=0, \tag{6.1}$$

where κ and δ are positive constants, $0<\varepsilon\leqq 1$ is a small parameter, and $w(t)$ is the Brownian motion process.

LEMMA 6.1.

$$E[|I(t)-I(s)|^4]\leqq 15\delta^4(t-s)^2\quad\textit{for } 0\leqq s\leqq t\quad\textit{uniformly in } 0\leqq\varepsilon\leqq 1. \tag{6.2}$$

Definition 6.1. For $t \geqq 0$ and $a > 0$, define the function $\lambda(t; a)$ by

$$\lambda(t; a) = \left(\frac{1}{a}\right)(1 - \exp[-at]).$$

Here we will prove (6.2) by showing the following estimates:

$$E[|I(t)|^2] = \delta^2 \lambda\left(t; \frac{2\kappa}{\varepsilon}\right) \quad \text{for } t \geqq 0, \tag{6.3}$$

$$E[I(u)I(s)] = \frac{1}{2}\delta^2\left[\lambda\left(u+s; \frac{\kappa}{\varepsilon}\right) - \lambda\left(u-s; \frac{\kappa}{\varepsilon}\right)\right] \quad \text{for } u \geqq s, \tag{6.3$'$}$$

$$E[|I(t)|^4] = 6\delta^4\left(\frac{\varepsilon}{2\kappa}\right)\left[\lambda\left(t; \frac{2\kappa}{\varepsilon}\right) - \lambda\left(t; \frac{4\kappa}{\varepsilon}\right)\right] \quad \text{for } t \geqq 0, \tag{6.4}$$

$$E[|I(t)|^4] \leqq \frac{3}{2}\left(\frac{\delta^4}{\kappa^2}\right) \times \varepsilon^2 \quad \text{for } t \geqq 0, \tag{6.4$'$}$$

$$\begin{aligned} &E[|I(t) - I(s)|^2] \\ &\quad = \delta^2\left[\left(1 - \exp\left[\frac{-2\kappa s}{\varepsilon}\right]\right)\right. \\ &\qquad \times \left\{\lambda\left(t-s; \frac{2\kappa}{\varepsilon}\right) + \lambda\left(t-s; \frac{\kappa}{\varepsilon}\right) - 2\lambda\left(t-s; \frac{2\kappa}{\varepsilon}\right)\right\} + \left.\lambda\left(t-s; \frac{2\kappa}{\varepsilon}\right)\right] \\ &\qquad\qquad \text{for } 0 \leqq s \leqq t, \end{aligned} \tag{6.5}$$

$$E[|I(t) - I(s)|^2] \leqq 3\delta^2(t-s) \quad \text{for } 0 \leqq s \leqq t, \tag{6.5$'$}$$

$$E[(I(u) - I(s))^3 I(s)] = -A(u; s) \quad \text{for } u \geqq s \tag{6.6}$$

with a nonnegative and continuous function $A(u; s)$ of $u \geqq s$ satisfying

$$\begin{aligned} A(u; s) &\leqq \frac{3}{2}\delta^4\left[\left(\frac{\varepsilon}{\kappa}\right)\lambda\left(u-s; \frac{\kappa}{\varepsilon}\right)\right. \\ &\qquad + \left.\left(1 - \exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right]\right)^2\left(\frac{\varepsilon}{\kappa}\right)\lambda\left(u-s; \frac{\kappa}{\varepsilon}\right)\right] \\ &\leqq 3\delta^4\left(\frac{\varepsilon}{\kappa}\right)(u-s) \quad \text{for } u \geqq s. \end{aligned} \tag{6.6$'$}$$

Step 1 (proof of (6.3) and (6.3)$'$). Since $I(u)$ has the form

$$I(u) = \exp\left[\frac{-\kappa u}{\varepsilon}\right]\int_0^u \exp\left[\frac{\kappa v}{\varepsilon}\right]\delta\, dw(v),$$

$$E[I(u)I(s)] = \delta^2\left(\frac{\varepsilon}{2\kappa}\right)\left[\exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right] - \exp\left[\frac{-\kappa(u+s)}{\varepsilon}\right]\right] \quad \text{for } u \geqq s.$$

Namely, (6.3) and (6.3)$'$ hold.

Step 2 (proof of (6.4) and (6.4)'). The Ito formula applies to (6.1) with the following result:

$$|I(t)|^4 = -\frac{4\kappa}{\varepsilon}\int_0^t |I(u)|^4\,du + 6\delta^2\int_0^t |I(u)|^2\,du$$
$$+4\delta\int_0^t I(u)^3\,dw(u).$$

As is already shown by (4.6), since

$$E[|I(t)|^{2m}] < \infty \quad \text{for all } t \geqq 0 \quad \text{with } m \geqq 1,$$

we can take expectation on the both sides of the above equation and get

$$E[|I(t)|^4] = -\frac{4\kappa}{\varepsilon}\int_0^t E[|I(u)|^4]\,du + 6\delta^2\int_0^t E[|I(u)|^2]\,du.$$

For $t \geqq 0$, define the function $f_{2m}(t)$ by

$$f_{2m}(t) = E[|I(t)|^{2m}], \quad \text{where } m \geqq 1.$$

Then $f_4(t)$ satisfies the linear ordinary differential equation of first order:

$$\frac{d}{dt}f_4(t) = -\frac{4\kappa}{\varepsilon}f_4(t) + 6\delta^2 f_2(t), \qquad f_4(0) = 0,$$

where $f_2(t) = E[|I(t)|^2]$ is given by (6.3). Therefore,

$$f_4(t) = \exp\left[\frac{-4\kappa t}{\varepsilon}\right]\int_0^t \exp\left[\frac{4\kappa u}{\varepsilon}\right]6\delta^2 \times \delta^2\lambda\left(u; \frac{2\kappa}{\varepsilon}\right)du$$
$$= 6\delta^4\left(\frac{\varepsilon}{2\kappa}\right)\exp\left[\frac{-4\kappa t}{\varepsilon}\right]$$
$$\times\left[\frac{\varepsilon}{4\kappa}\left(\exp\left[\frac{4\kappa t}{\varepsilon}\right]-1\right) - \frac{\varepsilon}{2\kappa}\left(\exp\left[\frac{2\kappa t}{\varepsilon}\right]-1\right)\right],$$

which yields (6.4). The estimate (6.4)' follows from (6.4) and the inequality that if $0 < a < b$, then

$$0 \leqq \lambda(t; a) - \lambda(t; b) \leqq \frac{1}{a} \quad \text{for } t \geqq 0.$$

Step 3 (proof of (6.5) and (6.5)'). By the Ito formula applied to (6.1), we see that for $0 \leqq s \leqq t$,

$$|I(t) - I(s)|^2$$
$$= -\frac{2\kappa}{\varepsilon}\int_s^t (I(u) - I(s))I(u)\,du$$
$$+\int_s^t \delta^2\,du + 2\int_s^t (I(u) - I(s))\delta\,dw(u)$$
$$= -\frac{2\kappa}{\varepsilon}\int_s^t |I(u) - I(s)|^2\,du - \frac{2\kappa}{\varepsilon}\int_s^t (I(u) - I(s))I(s)\,du$$
$$+\int_s^t \delta^2\,du + 2\int_s^t (I(u) - I(s))\delta\,dw(u).$$

Take expectation on the above equation, which is possible by the estimate (4.6). For $t \geqq s$, define the function $f_{2m}(t;\, s)$ by

$$f_{2m}(t;\, s) = E[|I(t) - I(s)|^{2m}], \quad \text{where } m \geqq 1.$$

Then $f_2(t;\, s)$ satisfies

$$f_2(t;\, s) = -\frac{2\kappa}{\varepsilon}\int_s^t f_2(u;\, s)\, du - \frac{2\kappa}{\varepsilon}\int_s^t E[(I(u) - I(s))I(s)]\, du$$
$$+ \int_s^t \delta^2\, du \quad \text{for } t \geqq s.$$

It follows from (6.3) and (6.3)′ that for $s \leqq u \leqq t$,

$$E[(I(u) - I(s))I(s)] = E[I(u)I(s)] - E[|I(s)|^2]$$
$$= -\delta^2\left(\frac{\varepsilon}{2\kappa}\right) a(u;\, s),$$

where

$$a(u;\, s) = \left(1 - \exp\left[\frac{-\kappa(u - s)}{\varepsilon}\right]\right)\left(1 - \exp\left[\frac{-2\kappa s}{\varepsilon}\right]\right).$$

Namely, $f_2(t;\, s)$ satisfies the linear ordinary differential equation of first order:

$$\frac{d}{dt} f_2(t;\, s) = -\frac{2\kappa}{\varepsilon} f_2(t;\, s) + \delta^2(a(t;\, s) + 1) \quad \text{for } t \geqq s,$$
$$f_2(s;\, s) = 0.$$

The solution is computed as follows:

$$f_2(t;\, s)$$
$$= \exp\left[\frac{-2\kappa(t - s)}{\varepsilon}\right]\int_s^t \exp\left[\frac{2\kappa(u - s)}{\varepsilon}\right]\delta^2(a(u;\, s) + 1)\, du$$
$$= \delta^2 \exp\left[\frac{-2\kappa(t - s)}{\varepsilon}\right]$$
$$\cdot\left[\left(1 - \exp\left[\frac{-2\kappa s}{\varepsilon}\right]\right)\int_s^t \left(\exp\left[\frac{2\kappa(u - s)}{\varepsilon}\right] - \exp\left[\frac{\kappa(u - s)}{\varepsilon}\right]\right) du\right.$$
$$\left. + \int_s^t \exp\left[\frac{2\kappa(u - s)}{\varepsilon}\right] du\right]$$
$$= \delta^2\left[\left(1 - \exp\left[\frac{-2\kappa s}{\varepsilon}\right]\right)\right.$$
$$\cdot\left\{\frac{\varepsilon}{2\kappa}\left(1 - \exp\left[\frac{-2\kappa(t - s)}{\varepsilon}\right]\right)\right.$$
$$\left. - \frac{\varepsilon}{\kappa}\left(\exp\left[\frac{-\kappa(t - s)}{\varepsilon}\right] - \exp\left[\frac{-2\kappa(t - s)}{\varepsilon}\right]\right)\right\}$$
$$\left. + \frac{\varepsilon}{2\kappa}\left(1 - \exp\left[\frac{-2\kappa(t - s)}{\varepsilon}\right]\right)\right].$$

Rearrangement of this equation yields (6.5). Since $0 \leqq \lambda(u;\, a) \leqq u$ for all $u \geqq 0$ and $a > 0$, it follows from (6.5) that

$$f_2(t;\, s) \leqq \delta^2[\{(t-s)+(t-s)\}+(t-s)] \quad \text{for } t \geqq s,$$

which implies (6.5)$'$.

Step 4 (proof of (6.6) and (6.6)$'$). For $u \geqq s$, set

$$\tilde{I}(u;\, s) = I(u) - \exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right] I(s).$$

Then, for $u \geqq s$,

$$I(u) - I(s) = \tilde{I}(u;\, s) - \left(1 - \exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right]\right) I(s)$$

and

$$\begin{aligned}(I(u)-I(s))^3 I(s) &= \tilde{I}(u;\, s)^3 I(s)\\ &\quad -3\left(1-\exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right]\right)\tilde{I}(u;\, s)^2 I(s)^2\\ &\quad +3\left(1-\exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right]\right)^2 \tilde{I}(u;\, s) I(s)^3\\ &\quad -\left(1-\exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right]\right)^3 I(s)^4.\end{aligned}$$

Note that $\tilde{I}(u;\, s)$ and $I(s)$ are independent for $u > s$, since the Brownian motion $w(v)$ is a process with independent increments. So, if m and n are positive integers, then

$$E[\tilde{I}(u;\, s)^m I(s)^n] = E[\tilde{I}(u;\, s)^m] E[I(s)^n] \quad \text{for } u > s.$$

Moreover, for any $u \geq s$,

$$E[\tilde{I}(u;\, s)] = E[I(s)] = 0.$$

Therefore, for $u > s$,

$$\begin{aligned}&E[(I(u)-I(s))^3 I(s)]\\ (6.7)\qquad &\quad = -3\left(1-\exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right]\right) E[\tilde{I}(u;\, s)^2] E[|I(s)|^2]\\ &\qquad -\left(1-\exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right]\right)^3 E[|I(s)|^4].\end{aligned}$$

We note that (6.7) holds with $u = s$, since

$$E[(I(s)-I(s))^3 I(s)] = 0.$$

On the other hand, for $u \geqq s$,

$$\begin{aligned}E[\tilde{I}(u;\, s)^2] &= E\left[\left\{I(u) - \exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right] I(s)\right\}^2\right]\\ &= E[|I(u)|^2] - 2\exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right] E[I(u) I(s)]\\ &\quad + \exp\left[\frac{-2\kappa(u-s)}{\varepsilon}\right] E[|I(s)|^2].\end{aligned}$$

For $u \geqq s$, define $\tilde{f}_2(u; s)$ and $f_I(u, s)$ by

$$\tilde{f}_2(u; s) = E[\tilde{I}(u; s)^2] \quad \text{and} \quad f_I(u, s) = E[I(u)I(s)].$$

Use the notation that $f_{2m}(u) = E[|I(u)|^{2m}]$ with $m \geqq 1$. Then, for $u \geqq s$

$$\tag{6.8} \begin{aligned} \tilde{f}_2(u; s) &= f_2(u) - 2\exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right] f_I(u, s) \\ &\quad + \exp\left[\frac{-2\kappa(u-s)}{\varepsilon}\right] f_2(s). \end{aligned}$$

For $u \geqq s$, set

$$\tag{6.9} \begin{aligned} A(u; s) &= 3\left(1 - \exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right]\right)\tilde{f}_2(u; s) f_2(s) \\ &\quad + \left(1 - \exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right]\right)^3 f_4(s). \end{aligned}$$

Then, (6.7) can be written as

$$E[(I(u) - I(s))^3 I(s)] = -A(u; s) \quad \text{for } u \geqq s.$$

By the definition of $\tilde{f}_2(u; s)$ and $f_{2m}(s)$, where $m = 1$ and 2, $A(u; s)$ is a nonnegative and continuous function of $u \geqq s$, and hence (6.6) holds.

Next we must evaluate $A(u; s)$ for $u \geqq s$. Since the estimates (6.3) and (6.3)′ hold,

$$f_2(u) \leqq \delta^2\left(\frac{\varepsilon}{2\kappa}\right) \quad \text{for } u \geqq 0 \quad \text{and} \quad f_I(u, s) \geqq 0 \quad \text{for } u \geqq s.$$

Thus, it follows from (6.8) that for $u \geqq s$,

$$\begin{aligned} 0 \leqq \tilde{f}_2(u; s) &\leqq f_2(u) + \exp\left[\frac{-2\kappa(u-s)}{\varepsilon}\right] f_2(s) \\ &\leqq \delta^2\left(\frac{\varepsilon}{2\kappa}\right) + \exp\left[\frac{-2\kappa(u-s)}{\varepsilon}\right]\delta^2\left(\frac{\varepsilon}{2\kappa}\right) \\ &\leqq \delta^2\left(\frac{\varepsilon}{\kappa}\right). \end{aligned}$$

Combining this, (6.3), and (6.4)′ with (6.9) we see that for $u \geqq s$,

$$\begin{aligned} 0 \leqq A(u; s) &\leqq 3\left(1 - \exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right]\right)\delta^2\left(\frac{\varepsilon}{\kappa}\right) f_2(s) \\ &\quad + \left(1 - \exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right]\right)^3 f_4(s) \\ &\leqq 3\left(1 - \exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right]\right)\delta^2\left(\frac{\varepsilon}{\kappa}\right) \times \delta^2\left(\frac{\varepsilon}{2\kappa}\right) \\ &\quad + \left(1 - \exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right]\right)^3 \times \frac{3}{2}\frac{\delta^4}{\kappa^2} \times \varepsilon^2 \\ &= \frac{3}{2}\delta^4\left[\left(\frac{\varepsilon}{\kappa}\right)\lambda\left(u - s; \frac{\kappa}{\varepsilon}\right)\right. \\ &\quad \left. + \left(1 - \exp\left[\frac{-\kappa(u-s)}{\varepsilon}\right]\right)^2\left(\frac{\varepsilon}{\kappa}\right)\lambda\left(u - s; \frac{\kappa}{\varepsilon}\right)\right]. \end{aligned}$$

Namely, the first inequality of (6.6)′ holds. Note that

$$\lambda(v;\, a) \leqq v \quad \text{for } v \geqq 0 \quad \text{and} \quad a>0$$

and that

$$0 \leqq (1-\exp[-x])^2 \leqq 1-\exp[-x] < 1 \quad \text{for } x \geqq 0.$$

Then, by the first inequality of (6.6)′ we get

$$\begin{aligned} A(u;\, s) &\leqq \frac{3}{2}\delta^4\left[\left(\frac{\varepsilon}{\kappa}\right)(u-s)+\left(\frac{\varepsilon}{\kappa}\right)(u-s)\right] \\ &= 3\delta^4\left(\frac{\varepsilon}{\kappa}\right)(u-s) \quad \text{for all } u \geqq s, \end{aligned}$$

showing the second inequality of (6.6)′.

Proof of Lemma 6.1. The Ito formula applied to (6.1) yields that for $0 \leqq s \leqq t$,

$$\begin{aligned} &|I(t)-I(s)|^4 \\ &= -\frac{4\kappa}{\varepsilon}\int_s^t (I(u)-I(s))^3 I(u)\, du \\ &\quad +6\delta^2\int_s^t |I(u)-I(s)|^2\, du + 4\delta\int_s^t (I(u)-I(s))^3\, dw(u) \\ &= -\frac{4\kappa}{\varepsilon}\int_s^t |I(u)-I(s)|^4\, du - \frac{4\kappa}{\varepsilon}\int_s^t (I(u)-I(s))^3 I(s)\, du \\ &\quad +6\delta^2\int_s^t |I(u)-I(s)|^2\, du + 4\delta\int_s^t (I(u)-I(s))^3\, dw(u). \end{aligned}$$

Take expectation on the above equation, which is possible by the estimate (4.6). Use the function

$$f_{2m}(t;\, s) = E[|I(t)-I(s)|^{2m}] \quad \text{for } t \geqq s, \quad \text{where } m \geqq 1.$$

Then, for $u \geqq s$,

$$\begin{aligned} f_4(t;\, s) &= -\frac{4\kappa}{\varepsilon}\int_s^t f_4(u;\, s)\, du - \frac{4\kappa}{\varepsilon}\int_s^t E[(I(u)-I(s))^3 I(s)]\, du \\ &\quad +6\delta^2\int_s^t f_2(u;\, s)\, du. \end{aligned}$$

Namely, $f_4(t;\, s)$ satisfies the linear ordinary differential equation of first order:

$$\begin{aligned} \frac{d}{dt}f_4(t;\, s) &= -\frac{4\kappa}{\varepsilon}f_4(t;\, s) - \frac{4\kappa}{\varepsilon}E[(I(t)-I(s))^3 I(s)] \\ &\quad +6\delta^2 f_2(t;\, s) \quad \text{for } t \geqq s \end{aligned}$$

with the initial condition $f_4(s;\, s)=0$. The solution is given by

$$\begin{aligned} &f_4(t;\, s) \\ (6.10) \qquad &= \exp\left[\frac{-4\kappa(t-s)}{\varepsilon}\right] \\ &\quad \cdot \int_s^t \exp\left[\frac{4\kappa(u-s)}{\varepsilon}\right]\left\{-\frac{4\kappa}{\varepsilon}E[(I(u)-I(s))^3 I(s)] + 6\delta^2 f_2(u;\, s)\right\} du. \end{aligned}$$

Here (6.6) and (6.6)′ imply

$$-\frac{4\kappa}{\varepsilon}E[(I(u)-I(s))^3I(s)]=\frac{4\kappa}{\varepsilon}A(u;s)$$
$$\leqq 12\delta^4(u-s) \quad \text{for } u\geqq s.$$

On the other hand, (6.5)′ implies

$$6\delta^2 f_2(u;s)\leqq 18\delta^4(u-s) \quad \text{for } u\geqq s.$$

Combining these inequalities with (6.10), we obtain

$$f_4(t;s)\leqq \exp\left[\frac{-4\kappa(t-s)}{\varepsilon}\right]\int_s^t \exp\left[\frac{4\kappa(u-s)}{\varepsilon}\right](30\delta^4(u-s))\,du$$
$$\leqq 30\delta^4\int_s^t (u-s)\,du$$
$$=15\delta^4(t-s)^2 \quad \text{for } t\geqq s.$$

Hence the proof of Lemma 6.1 is complete.

Acknowledgments. The author is grateful to the referee for his valuable advice, and to Professor Z. Schuss for his encouraging criticism.

REFERENCES

[1] J. GRASMAN, *Asymptotic Methods for Relaxation Oscillations and Applications*, Springer-Verlag, New York, 1987.

[2] R. Z. HASMINSKII, *Stochastic Stability of Differential Equations*, Sijthoff & Noordhoff, Alphen aan den Rijn, the Netherlands, 1980.

[3] A. H. NAYFEH, *Perturbation Methods*, John Wiley, New York, 1973.

[4] K. NARITA, *Remarks on nonexplosion theorem for stochastic differential equations*, Kodai Math. J., 5 (1982), pp. 395-401.

[5] G. C. PAPANICOLAOU, *Some probabilistic problems and methods in singular perturbations*, Rocky Mountain J. Math., 6 (1976), pp. 653-674.

[6] YU. V. PROKHOROV, *Convergence of random processes and limit theorems in probability theory*, Theory Probab. Appl., 1 (1956), pp. 157-214.

[7] A. V. SKOROKHOD, *Limit theorems for stochastic processes*, Theory Probab. Appl., 1 (1956), pp. 261-290.

[8] B. VAN DER POL, *Über Relaxations Schwingungen*, Jahrb. Drahtl. Telegr. Teleph., 28 (1927), pp. 178-184.

SIAM J. MATH. ANAL.
Vol. 24, No. 1, pp. 200–215, January 1993

SOLVABILITY OF BOUNDARY VALUE PROBLEMS FOR SYSTEMS OF SINGULAR DIFFERENTIAL-ALGEBRAIC EQUATIONS*

ROSWITHA MÄRZ[†] AND EWA B. WEINMÜLLER[‡]

This paper is dedicated to Professor Hans J. Stetter on the occasion of his 60th birthday.

Abstract. This paper considers systems of linear singular differential-algebraic equations subject to two-point boundary conditions. Existence and uniqueness theory is given for the case when one can decouple the original problem in a boundary value problem for ODEs with a singularity of the first kind at the left endpoint of the interval, and a set of algebraic equations. This paper also studies the solvability of the system in its initial value version.

Key words. differential-algebraic equations, singular systems of ODEs, existence and uniqueness of bounded solutions, projection methods

AMS(MOS) subject classifications. 34A10, 34B05, 34C20, 34C30

1. Introduction. Implicit singular ordinary differential equations (ODEs) result often from modelling of phenomena in applied sciences and technics. Therefore, recently, their analytical properties essential for an efficient numerical treatment have been extensively studied. Typically, linear systems are of the form

$$\tilde{A}(t)x'(t) + \tilde{B}(t)x(t) = \tilde{f}(t), \qquad t \in (0,1], \tag{1.1}$$

where the coefficient matrix $\tilde{A}(t)$ is singular.

In numerous articles on analysis and numerical solution of linear and nonlinear differential-algebraic equations (DAEs) existence of sufficiently smooth solutions is assumed. Consequently, the Taylor series expansion of such a solution provides a tool in the study of the analytical features of the problem, and the properties of numerical methods applied for its approximate solution. We refer to [3] and [9] for more detailed information on this technique. Obviously, this approach justifies the question, under which circumstances can we guarantee the required smoothness of the solution $x(t)$ of (1.1). The answer in the case of constant coefficient matrices $\tilde{A}$ and $\tilde{B}$ is easy: smoothness of x depends merely on the smoothness of the right-hand side $\tilde{f}$. It is more complicated for the general problem (1.1); we have to consider additionally the smoothness of certain canonical subspaces; cf. [1] and [8]. One of the most important is the nullspace of $\tilde{A}(t)$. As defined in [1], we call the system of DAEs "normal" if rank of $\tilde{A}(t)$ is constant, and "singular" if it is not. Here, all phenomena known from classical theory of ODEs with singularities may occur.

The simplest class of normal DAEs are the so-called "transferable," or "index 1" systems; see [1] and [7]. Essentially, their solution $x(t)$ can be split into two parts, $x(t) = u(t) + v(t)$, such that $u(t)$ solves a set of classical ODEs, subject to boundary or initial conditions, and $v(t)$ satisfies a set of algebraic equations. Moreover, it can be shown that "index 1" problems are well posed and can be solved numerically in, more or less, a standard manner. Here, we study systems (1.1), where $\tilde{A}(t) \equiv tA(t)$,

$$tA(t)x'(t) + B(t)x(t) = \tilde{f}(t), \qquad t \in (0,1], \tag{1.2}$$

[*] Received by the editors December 11, 1989; accepted for publication (in revised form) May 5, 1992.

[†] Humboldt-Universität zu Berlin, Sektion Mathematik, Unter den Linden 6, DO-1086 Berlin, Germany.

[‡] Institut für Angewandte und Numerische Mathematik, Technische Universität Wien, Wiedner Hauptstrasse 8-10/115, A-1040 Wien, Austria.

and the rank of $A(t)$ is constant as a function of t. This means that the system is transferable on $(0,1]$, and there is a rank change in matrix $\tilde{A}(t)$ for $t=0$. We show the solvability statements for initial and boundary value problems. It turns out that the eigenvalue stucture of the pair $\{A(0),\ B(0)\}$ already determines the first derivative of $x(t)$ and, in general, $x'(0)$ may become unbounded. This is a well-known fact in the theory of singular systems of ODEs; cf. [2], [5], and [6]. Thus a theory where the assumption of arbitrarily smooth solution remains unrelated to the rank of $\tilde{A}(t)$ seems questionable.

The following investigation of (1.2) may serve as basic knowledge in the further study of more involved singular DAEs, but it is also motivated by applications, where singular boundary value problems for systems of ODEs are augmented by sets of algebraic constraints. We derive the results combining techniques from the theory of transferable DAEs and the theory of singular boundary value problems. The crucial common element in both concepts is the use of projection matrices, which, appropriately chosen, guarantee certain commutativity properties, stated in Lemmas 2.2, 2.3, and 2.4. We can successfully link both techniques by partitioning the original boundary value problem for a system of DAEs into an equivalent system of singular boundary value problem of ODEs and a set of algebraic equations; see [4] and [7]. This equivalence is shown in Theorems 2.1 and 3.2. The paper is organized as follows: in §2, we study the constant coefficient case. We define the projection matrices, use them to construct a continuous solution $x \in C[0,1]$, and formulate conditions necessary for its uniqueness. In §3, we extend the existence and uniqueness results to the variable coefficient case. In particular, we heavily rely on the construction of the general solution presented in §2.

2. Constant coefficient case. We investigate the linear boundary value problem

$$tAx'(t)+Bx(t)=tf(t), \qquad 0<t\leq 1, \tag{2.1a}$$

$$B_0x(0)+B_1x(1)=\beta, \tag{2.1b}$$

where A, B are constant real $m\times m$ matrices, B_0, B_1 are constant real $k\times m$ matrices, such that $\operatorname{rank}[B_0B_1]=k, k\leq m$, and $f\in C[0,1]$ is a vector-valued function of dimension m. Here, we assume A to be singular, and denote by $Q\in \mathbb{R}^{m\times m}$ a projection matrix on the kernel of A, $\ker(A)=: N_0$. Finally, with $P:=I-Q$, we can reformulate (2.1a) and obtain

$$A(tPx(t))'+(B-A)x(t)=tf(t), \qquad 0<t\leq 1. \tag{2.2}$$

This suggests restricting the space of solutions x of (2.2) to

$$C^1_\pi := \{x\in C[0,1]:\ \pi x\in C^1[0,1],\ (\pi x)(t):=tPx(t),\ t\in[0,1]\}.$$

We note that $C^1_\pi \not\subseteq C^1[0,1]$, in general. Moreover, C^1_π does not depend on the special choice of Q, but it is entirely determined by $\ker(A)$. For the convenience of notation we use $C:=C[0,1]$, $C^1:=C^1[0,1]$, and $C^1_0:=C\cap C^1(0,1]$. Before discussing the system in its general form (2.1a), we consider an example in order to present the basic idea of the decoupling technique.

2.1. Case study. Given the problem

$$tAx'(t)+Bx(t)=tf(t), \qquad 0<t\leq 0, \tag{2.3}$$

where $m = 3$, $x = (x_1, x_2, x_3)^{\mathrm{T}}$ and

$$A = \operatorname{diag}(1,1,0), \quad B = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad f = (f_1, f_2, f_3)^{\mathrm{T}}.$$

Since (2.3) does not include x_3', we attempt to decouple the system into two differential equations for $(x_1, x_2)^{\mathrm{T}}$ and one algebraic equation for x_3. Some elementary calculations yield,

$$t\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} x_1'(t) \\ x_2'(t) \\ 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ x_3(t) \end{pmatrix}$$
$$+\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{pmatrix}\begin{pmatrix} x_1(t) \\ x_2(t) \\ 0 \end{pmatrix} = t\begin{pmatrix} f_1(t) \\ f_2(t) \\ f_3(t) \end{pmatrix},$$

or equivalently, with $Q = \operatorname{diag}(0,0,1)$ and $P = \operatorname{diag}(1,1,0)$,

$$tGu'(t) + Gv(t) + BPu(t) = tf(t), \qquad 0 < t \leq 1, \tag{2.4}$$

where

$$x(t) \equiv Px(t) + Qx(t) =: u(t) + v(t), \qquad G = A + BQ.$$

Since G is regular, we multiply (2.4) by G^{-1} and have

$$tu'(t) + v(t) + \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} u(t) = t\begin{pmatrix} f_1(t) - 2f_3(t) \\ f_2(t) \\ 2f_3(t) \end{pmatrix},$$

which provides the desired decomposition,

$$\begin{pmatrix} x_1'(t) \\ x_2'(t) \end{pmatrix} = \frac{1}{t}\begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} + \begin{pmatrix} f_1(t) - 2f_2(t) \\ f_2(t) \end{pmatrix}, \quad 0 < t \leq 1, \tag{2.5a}$$

$$x_3(t) = -x_2(t) + 2tf_3(t). \tag{2.5b}$$

Clearly, the above separation method becomes impracticable if G is singular, so the crucial assumption for the investigation of the general case is the *regularity* of the matrix

$$G = A + BQ \equiv AP + BQ.$$

2.2. General case. We now execute the procedure presented in §2.1 to transform the general problem (2.1) and study its properties. We assume that the system (2.1a) is transferable for $t \in (0,1]$, or equivalently, $G = A + BQ$ is nonsingular; cf. [4]. It has been shown in [1, Thm. A.13] that two matrices A and B form a regular matrix pencil of index 1 if and only if A is singular, but G is not. Moreover, if G is nonsingular, then $Q_s := QG^{-1}B$ represents the projection onto N_0, along the subspace

$$S := \{z \in \mathbb{R}^m : Bz \in \operatorname{im}(A)\}.$$

We also denote $P_s := I - Q_s$.

THEOREM 2.1. *Let $x \in C^1_\pi$ satisfy the boundary value problem* (2.1). *Then the function pair $u := Px \in C^1_0$, $v := Qx \in C$ is a solution of the system*

$$tu'(t) + PG^{-1}Bu(t) = tPG^{-1}f(t), \qquad 0 < t \le 1, \tag{2.6a}$$

$$B_0P_su(0) + B_1P_su(1) = \beta - B_1QG^{-1}f(1), \tag{2.6b}$$

$$Qu(\delta) = 0, \tag{2.6c}$$

$$v(t) = -Q_su(t) + tQG^{-1}f(t), \qquad 0 < t \le 1, \tag{2.6d}$$

where $\delta \in (0, 1]$ is arbitrary but fixed. If $u \in C^1_0$, $v \in C$ solve the transformed system (2.6), *then $x := u + v \in C^1_\pi$ is a solution of the original problem* (2.1).

Proof. It is easy to see that $G^{-1}A = P$ and $G^{-1}BQ = Q$. Using these relations, we can reformulate (2.1a),

$$t(Px)'(t) + Qx(t) + G^{-1}BPx(t) = tG^{-1}f(t),$$

which implies

$$t(Px)'(t) + PG^{-1}BPx(t) = tPG^{-1}f(t), \tag{2.7a}$$

$$Qx(t) + Q_sPx(t) = tQG^{-1}f(t). \tag{2.7b}$$

Consequently, any solution $x \in C^1_\pi$ of (2.1a) solves (2.6a), (2.6d). Trivially, (2.1b) can be written as

$$B_0(u(0) + v(0)) + B_1(u(1) + v(1)) = \beta. \tag{2.8}$$

We evaluate (2.6d) at $t = 0$, and $t = 1$, substitute $v(0)$, $v(1)$ into (2.8), and obtain (2.6b),

$$B_0\underbrace{(I - Q_s)}_{P_s}u(0) + B_1\underbrace{(I - Q_s)}_{P_s}u(1) = \beta - B_0QG^{-1}f(1).$$

Finally, (2.6c) follows from $Q(Px(\delta)) = 0$, for any δ, and this concludes the proof of the first statement. Let us assume that $u \in C^1_0$, $v \in C$ solve the system (2.6). Then the multiplication of (2.6a) by Q and (2.6c) imply

$$t(Qu)'(t) = 0 \Rightarrow Qu(t) = \text{constant},$$

and since $Qu(\delta) = 0$, we have

$$Qu(t) = 0 \Rightarrow Pu(t) = u(t).$$

From the multiplication of (2.6d) by Q we conclude $v(t) = Qv(t)$. Consequently, $x := u + v \equiv Pu + Qv \in C^1_\pi$, and the result follows. □

According to Theorem 2.1, we can characterize solvability of (2.1) via discussion of (2.6a), (2.6b), and (2.6c). In the theory of singular boundary value problems, the conventional form of (2.6a) is

$$u'(t) = \frac{1}{t}Mu(t) + F(t), \qquad 0 < t \le 1, \tag{2.9}$$

where $M := -PG^{-1}B$ and $F(t) := PG^{-1}f(t)$. We use techniques developed in [2] to select a bounded solution $u \in C^1_0$ of (2.9).

Let us denote by $\Phi(t) \in C^1(0,1]$ the fundamental solution matrix of the homogeneous problem

$$\Phi'(t) = \frac{1}{t}M\Phi(t), \quad 0 < t \leq 1, \qquad \Phi(\delta) = I. \tag{2.10}$$

Then,

$$\Phi(t) = \exp\left(M \ln \frac{t}{\delta}\right) =: \left(\frac{t}{\delta}\right)^M,$$

and the general solution of (2.9) is

$$u(t) = \Phi(t)\left\{c + \int_\delta^t \Phi^{-1}(s)F(s)ds\right\}, \qquad 0 < t \leq 1. \tag{2.11}$$

It follows immediately that $u \in C[\delta, 1]$, but in general, $u \notin C[0,1]$.

LEMMA 2.2.

$$\Phi(t)P = P\Phi(t), \qquad \Phi(t)Q = Q\Phi(t) = Q \quad \forall t \in (0,1].$$

Proof.

$$MQ = \underbrace{-PG^{-1}B}_{M}Q = -P\underbrace{G^{-1}BQ}_{Q} = -PQ = 0 = Q(\underbrace{-PG^{-1}B}_{M}) = QM$$

$$\Rightarrow \Phi(t)Q = \left(I + \sum_{i=1}^{\infty}\frac{1}{i!}(tM)^i\right)Q = Q\Phi(t) = Q.$$

Moreover,

$$MP = M(P+Q) = M = -PG^{-1}B = -P^2G^{-1}B = PM \quad \Rightarrow \quad \Phi(t)P = P\Phi(t). \quad \square$$

Remark. $\Phi(t)P$ and $P\Phi(t)$ are identical solutions of (2.10), with $\Phi(\delta)P = P$. To show $Q\Phi(t) = Q$, we could also argue as follows:

$$(Q\Phi(t))' = Q\Phi'(t) = \frac{1}{t}QM\Phi(t) = -\frac{1}{t}QP\cdots \equiv 0 \Rightarrow Q\Phi(t) = \text{constant} = Q.$$

Using the commutativity rules given by Lemma 2.2, and $F(s) \equiv PF(s)$, $Q\Phi^{-1}(s)P \equiv 0$, we can rewrite (2.11), and obtain a new form for the general solution $u(t)$,

$$u(t) = Qc + \Phi(t)P\left\{c + \int_\delta^t \Phi^{-1}(s)F(s)ds\right\}, \qquad 0 < t \leq 1. \tag{2.12}$$

Obviously, condition (2.6c) holds if and only if

$$Qc = 0, \tag{2.13}$$

and any $u(t)$ given by (2.12) and satisfying (2.13) belongs to $\operatorname{im}(P)$.

So far, we have represented $\mathbb{R}^m$ as $\mathbb{R}^m = N_0 \oplus \operatorname{im}(P)$, with $N_0 \subseteq \ker(M)$. Now, we additionally decompose $\ker(M) := N_0 \oplus N_1$, in such a way that $N_1 \subseteq \operatorname{im}(P)$.

Moreover, we introduce the following notation:

X_0 is the invariant subspace of M, associated with eigenvalues $\lambda = 0$, in N_1;

X_- is the invariant subspace of M, associated with $\lambda = \sigma + i\eta \neq 0$ such that $\sigma \leq 0$;

X_+ is the invariant subspace of M, associated with $\lambda = \sigma + i\eta$, $\sigma > 0$.

Now, we immediately have

$$\mathbb{R}^m = N_0 \oplus X_0 \oplus X_- \oplus X_+, \qquad N_1 \subseteq X_0, \quad N_1 \subseteq \operatorname{im}(P). \tag{2.14}$$

LEMMA 2.3. *$X_0 \subseteq \operatorname{im}(P)$ can be chosen in such a way that $X_0 \oplus X_- \oplus X_+ = \operatorname{im}(P)$, holds.*

Proof. For a nontrivial eigenvalue $\lambda \neq 0$, we have

$$Mz_0 = \lambda z_0 \Longleftrightarrow PMz_0 = \lambda z_0 \Rightarrow Qz_0 = 0 \Longleftrightarrow Pz_0 = z_0.$$

Similarly, for any principal vector z_1, the defining equation $Mz_1 = \lambda z_1 + z_0$ implies $z_1 \in \operatorname{im}(P)$, and consequently, $X_- \oplus X_+ \subseteq \operatorname{im}(P)$. Let $\lambda = 0$ be an eigenvalue with the algebraic multiplicity $n_a = l+1$, and the geometric multiplicity $n_g = 1$. Then the set of generalized eigenvalues associated with λ solves

$$Mz_0 = 0, \quad Mz_1 = z_0, \ldots, \quad Mz_l = z_{l-1}, \quad z_0 \in N_1 \subseteq \operatorname{im}(P).$$

Since the general solution of $Mz_l = z_{l-1}$ is $z_l = \tilde{z}_l + \ker(M)$, where $M\tilde{z}_l = z_{l-1}$, z_l depends on $\dim(\ker(M))$ parameters. In general, $\dim(\ker(M)) \geq \dim(\ker(A)) = \operatorname{rank}(Q)$, and, therefore, the parameters can be fixed in such a way that $Qz_l = 0 \Leftrightarrow Pz_l = z_l$. For $k = l, l-1, \ldots, 2$,

$$Qz_{k-1} = QMz_k = 0 \Rightarrow z_{k-1} \in \operatorname{im}(P),$$

and the result follows. □

For the new decomposition of $\mathbb{R}^m$, given by (2.14), with $\operatorname{im}(P) = X_0 \oplus X_- \oplus X_+$, we need two additional projections $P_1, P_2 \in \mathbb{R}^{m\times m}$, such that

P_1 is a projection onto $X_0 \oplus X_-$, along the subspace $N_0 \oplus X_+$;

P_2 is a projection onto X_+, along the subspace $N_0 \oplus X_0 \oplus X_-$.

According to the construction,

$$P = P_1 + P_2, \qquad P_i = PP_i = P_iP, \quad i = 1, 2. \tag{2.15}$$

LEMMA 2.4.

$$\Phi(t)P_i = P_i\Phi(t), \qquad 0 < t \leq 1, \quad i = 1, 2.$$

Proof. Let $J = \operatorname{diag}(0, J_0, J_-, J_+)$ denote the Jordan canonical form of $M = EJE^{-1}$, where E is the matrix containing the generalized eigenvectors of M. Then we can write P as

$$E\begin{pmatrix} 0 & K_1 & K_2 & K_3 \\ 0 & I_0 & 0 & 0 \\ 0 & 0 & I_- & 0 \\ 0 & 0 & 0 & I_+ \end{pmatrix}E^{-1}.$$

From $MP = PM$, it follows immediately that $K_1J_0 = 0$, $K_2J_- = 0$,and $K_3J_+ = 0$ must hold and since J_-, J_+ are regular, $K_2 = K_3 = 0$. Consequently,

$$P_1 = E\begin{pmatrix} 0 & K_1 & 0 & 0 \\ 0 & I_0 & 0 & 0 \\ 0 & 0 & I_- & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}E^{-1}, \qquad P_2 = E\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_+ \end{pmatrix}E^{-1},$$

and $MP_i = P_iM$, $i = 1, 2$, which completes the proof. □

We now return to the formula (2.12), and use the projections P_1 and P_2 to derive the following, more precise representation for the general solution $u(t)$ of (2.9), subject to the initial condition (2.13),

$$\begin{aligned} u(t) &= \Phi(t)P_1c + \Phi(t)\int_\delta^0 P_1\Phi^{-1}(s)F(s)ds + \Phi(t)\int_0^t P_1\Phi^{-1}(s)F(s)ds \\ &\quad + \Phi(t)\left\{P_2c + \int_\delta^t P_2\Phi^{-1}(s)F(s)ds\right\}, \end{aligned}$$

which yields

$$\begin{aligned} u(t) &= \Phi(t)\left\{P_1c - \int_0^\delta P_1\Phi^{-1}(s)F(s)ds\right\} + t\int_0^1 P_1\Phi^{-1}(s)\Phi(1)F(st)ds \\ &\quad + \Phi(t)\left\{P_2c + \int_\delta^t P_2\Phi^{-1}(s)F(s)ds\right\}, \qquad 0 < t \leq 1. \end{aligned} \tag{2.16}$$

It is clear from the form of (2.16) that the behavior of $u(t)$, in particular for $t \to 0$, depends essentially on the structure of $\Phi(t)$. We first note that

$$\begin{aligned} \Phi(t)P_1 &= E \ \operatorname{diag}\left(0, \left(\frac{t}{\delta}\right)^{J_0}, \left(\frac{t}{\delta}\right)^{J_-}, 0\right) E^{-1}, \\ \Phi(t)P_2 &= E \ \operatorname{diag}\left(0, 0, 0, \left(\frac{t}{\delta}\right)^{J_+}\right) E^{-1}. \end{aligned}$$

For any Jordan box

$$L := \begin{pmatrix} \lambda & 1 & & & 0 \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ 0 & & & & \lambda \end{pmatrix}, \qquad L \in \mathbb{C}^{n\times n},$$

we deduct from the definition of $\Phi(t)$,

$$t^L = t^\lambda \begin{pmatrix} 1 & \ln t & \dots & \dots & \frac{(\ln t)^{n-1}}{(n-1)!} \\ & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & \ln t \\ & & & & 1 \end{pmatrix}, \qquad t > 0,$$

and, consequently,

$$\lim_{t\to 0} \Phi(t)P_2 = 0,$$

$$\lim_{t\to 0} \Phi(t)\int_\delta^t P_2\Phi^{-1}(s)F(s)ds = 0.$$

Moreover, these parts of $\Phi(t)P_1$, contributed by the invariant subspaces associated with $\lambda = \sigma + i\eta \neq 0$, $\sigma \leq 0$, and by subspaces spanned by the principal vectors associated with $\lambda = 0$, have no limit, or become unbounded for $t \to 0$. Since the subspace spanned by the eigenvectors to $\lambda = 0$ provides bounded contribution in $\Phi(t)P_1$, we have to split X_0 again in order to separate the unbounded and bounded parts. Let $X_0 := N_1 \oplus H_1$, where H_1 is spanned by the principal vectors related to $\lambda = 0$. Then (2.14) becomes

$$\mathbb{R}^m = N_0 \oplus N_1 \oplus \underbrace{H_1 \oplus X_-}\oplus X_+. \tag{2.17}$$

We denote by R, the projection onto N_1, along $N_0 \oplus H_1 \oplus X_- \oplus X_+$, and by $H := P_1 - R$, the projection onto $H_1 \oplus X_-$, along $N_0 \oplus N_1 \oplus X_+$.

Then, $H = HP_1 = P_1H$, and

$$\Phi(t)R = R, \qquad 0 < t \leq 1.$$

We require,

$$Hc = \int_0^\delta H\Phi^{-1}(s)F(s)ds \Rightarrow Hu(0) = 0, \tag{2.18}$$

and obtain from (2.16),

$$\begin{aligned} u(t) &= Rc - \int_0^\delta R\Phi^{-1}(s)F(s)ds + t\int_0^1 P_1\Phi^{-1}(s)\Phi(1)F(st)ds \\ &+\Phi(t)\left\{P_2c + \int_\delta^t P_2\Phi^{-1}(s)F(s)ds\right\}, \qquad 0 < t \leq 1. \end{aligned} \tag{2.19}$$

This immediately implies

$$\lim_{t\to 0} u(t) = Rc - \int_0^\delta R\Phi^{-1}(s)F(s)ds, \tag{2.20}$$

and, therefore, (2.18) is necessary for $u \in C_0^1$. Finally, we consider the boundary conditions (2.6b) in order to determine the constants in P_2c and Rc. From the evaluation of (2.19) for $t = 0$ and $t = 1$, and substitution into (2.6b), we have

$$B_0P_sRc + B_1P_s(Rc + \Phi(1)P_2c) = \tilde{\beta}, \tag{2.21}$$

where

$$\begin{aligned} \tilde{\beta} := \beta - B_1QG^{-1}f(1) + B_0P_s\int_0^\delta R\Phi^{-1}(s)F(s)ds \\ - B_1P_s\left\{ - \int_0^\delta R\Phi^{-1}(s)F(s)ds + \int_0^1 P_1\Phi^{-1}(s)\Phi(1)F(s)ds \right. \\ \left. + \Phi(1)\int_\delta^1 P_2\Phi^{-1}(s)F(s)ds\right\}, \end{aligned}$$

and so, the boundary value problem (2.6) is uniquely solvable if the linear system (2.21) can be uniquely solved for the remaining free components of c. This is the case when for the coefficient matrix

$$T := B_0P_sR + B_1P_s\Phi(1)(R + P_2), \qquad T \in \mathbb{R}^{k\times m}, \tag{2.22}$$

the following conditions hold:

$$\operatorname{rank}(T) = k = \dim(N_1) + \dim(X_+), \tag{2.23}$$

$$\ker(T) = \ker(R + P_2) = N_0 \oplus H_1 \oplus X_-. \tag{2.24}$$

We summarize results from this section in the following theorem.

THEOREM 2.5. *The boundary value problem* (2.1) *has a unique solution* $x \in C^1_\pi$, *for arbitrary* $f \in C$, *and* $\beta \in \mathbb{R}^k$ *if and only if the coefficient matrix* T *given by* (2.22) *satisfies* (2.23) *and* (2.24).

COROLLARY 2.6. *The initial value problem* (2.1), $B_1 = 0$, *is uniquely solvable if and only if*

$$\Lambda := \{\lambda \in \mathbb{C}\backslash\{0\} : \det(\lambda A + B) = 0\} \subseteq \mathbb{C}_-, \tag{2.25}$$

and the following conditions hold for $T := B_0 P_s R$,

$$\begin{aligned}\operatorname{rank}(T) &= k = \dim(N_1),\\ \ker(T) &= \ker(R) = N_0 \oplus H_1 \oplus X_-.\end{aligned}$$

Proof. We first show that Λ and the set of nontrivial eigenvalues of M coincide. Let $\lambda \neq 0$ be an eigenvalue of M; then $z = Pz$ follows from $Mz = \lambda z$. Also,

$$\begin{aligned}
&-Mz + \lambda z = PG^{-1}BPz + \lambda Pz = 0\\
&\quad\Longleftrightarrow (I - Q)G^{-1}BPz + \lambda Pz = G^{-1}BPz - \underbrace{QG^{-1}B}_{Q_s} Pz + \lambda Pz = 0\\
&\quad\Longleftrightarrow BPz - GQ_sPz + \lambda\underbrace{(A + BQ)}_{G}Pz = BPz - GQ_sPz + \lambda APz = 0\\
&\quad\Longleftrightarrow \lambda APz + BPz - \underbrace{(A + BQ)}_{G}\underbrace{QG^{-1}B}_{Q_s} Pz = \lambda APz + BPz - BQ_sPz = 0\\
&\quad\Longleftrightarrow \lambda AP_sPz + B\underbrace{(I - Q_s)}_{P_s}Pz = (\lambda A + B)P_sz = 0.
\end{aligned}$$

Consequently, $\det(\lambda A + B) = 0$ and $\lambda \in \Lambda$.

Let us now assume $\lambda \in \Lambda$; then there exists a vector w such that

$$(\lambda A + B)w = 0, \qquad w \neq 0. \tag{2.26}$$

Multiplying (2.26) by PG^{-1}, we immediately obtain

$$\lambda P\underbrace{G^{-1}A}_{P}w + \underbrace{PG^{-1}B}_{-M=-MP}\, w = 0 \quad\Longleftrightarrow\quad (\lambda I - M)Pw = 0.$$

Moreover, multiplication of (2.26) by QG^{-1}, yields

$$\begin{aligned}
&\lambda Q\underbrace{G^{-1}A}_{P}w + \underbrace{QG^{-1}B}_{Q_s}(Qw + Pw) = 0\\
&\Longleftrightarrow Q\underbrace{G^{-1}BQ}_{I-P}w = -Q_sPw \quad\Longleftrightarrow\quad Qw = -Q_sPw.
\end{aligned}$$

This completes the proof, since $w \neq 0$ implies $Pw \neq 0$, and, therefore, $\det(\lambda I - M) = 0$ must hold. We conclude that Λ is the set of nontrivial eigenvalues of M, and it follows from $\Lambda \subseteq \mathbb{C}_-$ that X_+ is an empty set and $P_2 = 0$. □

Remarks.

1. If $X_0 \oplus X_+ = \emptyset$, then $\ker(M) = N_0$, and $\mathbb{R}^m = N_0 \oplus X_-$. In this case P_1 is a projection onto $X_- = \operatorname{im}(P), P = P_1 + P_2 = P_1 = R + H = H$. Also, $T = 0$, and the general solution (2.19) contains no free constants. It is $\lim_{t\to 0} u(t) = 0$, and $u \in C_0^1$, if it is subject to the homogeneous initial condition $u(0) = 0$. If the general solution given by (2.16) does not satisfy (2.18), then it becomes unbounded or has no limit as $t \to 0$.

2. If the smallest positive real part of the eigenvalues of M is sufficiently large, namely, $\sigma_{\min} > 1$, then $u \in C$ implies $u \in C^1$. This is also the case when, for the eigenvalue $\lambda \equiv \sigma = 1$, its algebraic and geometric multiplicities coincide. Additionally, if $QG^{-1}f \in C_0^1$, then the solution of the boundary value problem (2.1), $x \in C^1$. This follows immediately from (2.6d). We stress that $QG^{-1}f \in C_0^1$ is *necessary* for x to be in C^1.

3. For $\delta = 1$, $\Phi(1) = I$ (cf. (2.10)) and according to (2.19), the general continuous solution of (2.6a) and (2.6c) is

$$\begin{aligned} u(t) &= Rc - \int_0^1 R\Phi^{-1}(s)F(s)ds + t\int_0^1 P_1\Phi^{-1}(s)F(st)ds \\ &+\Phi(t)\left\{P_2c + \int_1^t P_2\Phi^{-1}(s)F(s)ds\right\}, \qquad 0 \le t \le 1. \end{aligned} \tag{2.27}$$

The linear system (2.21) can then be simplified to

$$B_0P_sRc + B_1P_s(Rc + P_2c) = \tilde{\beta}, \tag{2.28}$$

where

$$\begin{aligned} \tilde{\beta} := \beta - B_1QG^{-1}f(1) + B_0P_s\int_0^1 R\Phi^{-1}(s)F(s)ds \\ -B_1P_s\int_0^1 H\Phi^{-1}(s)F(s)ds. \end{aligned}$$

3. Variable coefficient case. Here, we discuss the linear system

$$tA(t)x'(t) + B(t)x(t) = tf(t), \qquad 0 < t \le 1, \tag{3.1a}$$

$$B_0x(0) + B_1x(1) = \beta, \tag{3.1b}$$

where A, B are real-valued continuous $m \times m$ matrices, $A(t) \neq 0$, and all other data is set as in (2.1). We assume that the kernel of A,

$$\ker(A(t)) =: N_0(t), \tag{3.2}$$

is nontrivial and smooth on $[0, 1]$, or equivalently, there exists a projection function $Q(t) \in \mathbb{R}^{m\times m}$, $Q \in C[0, 1]$, which maps $\mathbb{R}^m$ pointwise onto $N_0(t)$. We also define, $P(t) := I - Q(t)$. Moreover, let

$$G(t) := A(t) + B(t)Q(t) \tag{3.3}$$

be regular for all $t \in (0, 1]$. Under these assumptions, the differential-algebraic system (3.1a) is transferable on (0,1]. Recall that $G(t)$ is nonsingular if and only if the matrices $A(t)$, $B(t)$ form a regular matrix pencil of index 1. While the matrix pencil

$\{A(t),\ B(t)\}$ is regular with index 1 for $0 \le t \le 1$, the pencil $\{tA(t),\ B(t)\}$ is regular with index 1 for $0 < t \le 1$, and becomes singular at $t = 0$ when $B(0)$ is singular. The nullspace of $tA(t)$ is neither smooth nor of constant dimension, which changes in $t = 0$ from $m - \operatorname{rank}(A(t))$ to m.

Since,

$$A(t) \equiv A(t)P(t), \qquad tP(t)x'(t) = (tP(t)x(t))' - tP'(t)x(t) - P(t)x(t),$$

we rewrite (3.1a) and obtain

$$A(t)\{(tP(t)x(t))' - P(t)x(t)\} + (B(t) - tA(t)P'(t))x(t) = tf(t), \qquad 0 < t \le 0. \tag{3.4}$$

As previously, the solution space is

$$C^1_\pi := \{x \in C[0,1]:\ \pi x \in C^1[0,1],\ (\pi x)(t) := tP(t)x(t),\ t \in [0,1]\}.$$

As a first step in the investigation of (3.1) we derive the related, decoupled system analogous to (2.6). Again, $Q_s(t) := Q(t)G^{-1}(t)B(t)$, $P_s(t) := I - Q_s(t)$, where $Q_s(t)$ is the pointwise projection onto $N_0(t)$ along the subspace

$$S(t) := \{z \in \mathbb{R}^m:\ B(t)z \in \operatorname{im}(A(t))\}.$$

From the multiplication of (3.1a) by $P(t)G^{-1}(t)$ and $Q(t)G^{-1}(t)$, we have

$$\begin{aligned} &tP(t)\underbrace{G^{-1}(t)A(t)}_{P(t)}x'(t) + P(t)G^{-1}(t)B(t)x(t) = tP(t)G^{-1}(t)f(t),\\ &\iff t(P(t)x(t))' - tP'(t)x(t) + P(t)G^{-1}(t)B(t)x(t) = tP(t)G^{-1}(t)f(t),\\ &\quad 0 < t \le 1, \end{aligned} \tag{3.5}$$

and

$$\begin{aligned} &tQ(t)\underbrace{G^{-1}(t)A(t)}_{P(t)}x'(t) + Q(t)G^{-1}(t)B(t)x(t) = tQ(t)G^{-1}(t)f(t),\\ &\iff Q(t)G^{-1}(t)B(t)x(t) = tQ(t)G^{-1}(t)f(t), \qquad 0 < t \le 1. \end{aligned} \tag{3.6}$$

We introduce $u := Px$ and $v := Qx$, and reformulate (3.6):

$$\begin{aligned} &\underbrace{Q(t)G^{-1}(t)B(t)}_{Q_s}u(t) + Q(t)\underbrace{G^{-1}(t)B(t)Q(t)}_{I-P(t)}x(t) = tQ(t)G^{-1}(t)f(t)\\ &\Rightarrow v(t) = -Q_s(t)u(t) + tQ(t)G^{-1}(t)f(t), \qquad 0 < t \le 1. \end{aligned} \tag{3.7}$$

This implies

$$x(t) = u(t) + v(t) = P_s(t)u(t) + tQ(t)G^{-1}(t)f(t), \tag{3.8}$$

and the substitution of (3.8) into (3.5) yields

$$tu'(t) - M(t)u(t) = tF(t), \qquad 0 < t \le 1, \tag{3.9}$$

where

$$M(t) := -P(t)G^{-1}(t)B(t) + tP'(t)P_s(t),$$
$$F(t) := P(t)G^{-1}(t)f(t) + tP'(t)Q(t)G^{-1}(t)f(t).$$

Due to (3.8), the boundary conditions (3.1b) read

$$B_0P_s(0)u(0) + B_1P_s(1)u(1) = \beta - B_1Q(1)G^{-1}(1)f(1). \tag{3.10}$$

In order to be able to combine solutions $u(t)$, $v(t)$ of (3.9), and (3.7) to a solution $x(t)$ of (3.1a), it is necessary that $u(t) \in \mathrm{im}(P(t))$, $t \in (0,1]$.

LEMMA 3.1. *Let $u(t) \in C_0^1$ be a solution of* (3.9), *subject to*

$$Q(\delta)u(\delta) = 0, \qquad \delta \in (0,1]. \tag{3.11}$$

Then $Q(t)u(t) = 0$ for all $t \in (0,1]$.

Proof. $QF = 0$ follows from $QP'Q = -QPQ' = 0$. We multiply (3.9) by Q and conclude

$$tQ(t)u'(t) - tQ(t)P'(t)P_s(t)u(t) = 0,$$

and finally,

$$t\{Q(t)u'(t) + Q'(t)P(t)u(t)\} = 0,$$
$$t\{(Q(t)u(t))' - Q'(t)Q(t)u(t)\} = 0.$$

Consequently, for $U(t) := Q(t)u(t)$, $t > 0$, we obtain the following homogeneous initial value problem with trivial initial condition

$$U'(t) - Q'(t)U(t) = 0, \qquad 0 < t \le 1,$$
$$U(\delta) = 0,$$

which implies $U(t) \equiv Q(t)u(t) = 0, 0 < t \le 1$, and the result follows. Clearly, $u(t) \in \mathrm{im}(P(t))$, $0 < t \le 1$. □

Let us unite (3.7), (3.9), (3.10), and (3.11) to a system

$$tu'(t) - M(t)u(t) = tF(t), \qquad 0 < t \le 1, \tag{3.12a}$$
$$B_0P_s(0)u(0) + B_1P_s(1)u(1) = \beta - B_1Q(1)G^{-1}(1)f(1), \tag{3.12b}$$
$$Q(\delta)u(\delta) = 0, \qquad \delta \in (0,1], \tag{3.12c}$$
$$v(t) = -Q_s(t)u(t) + tQ(t)G^{-1}(t)f(t), \qquad 0 < t \le 1, \tag{3.12d}$$

and formulate the following assertion, in analogy to Theorem 2.1.

THEOREM 3.2. *Let $x \in C_\pi^1$ be a solution of* (3.1); *then the pair, $u := Px \in C_0^1$, $v := Qx \in C$ satisfies* (3.12). *For any $u \in C_0^1$ and $v \in C$, solving the system* (3.12), *we construct the solution of the boundary value problem* (3.1) *by $x := u + v \in C_\pi^1$.*

In order to find continuous solutions of (3.12), we first study the solution manifold of the singular system (3.12a),

$$u'(t) = \frac{1}{t}M(t)u(t) + F(t), \qquad 0 < t \le 1. \tag{3.13}$$

Let us assume that $M'(0)$ exists; then the matrix $M_1(t)$,

(3.14)
$$M_1(t) := \frac{1}{t}(M(t) - M(0)) = -\frac{1}{t}(P(t)G^{-1}(t)B(t) - P(0)G^{-1}(0)B(0)) + P'(t)P_s(t),$$

is continuous on $[0,1]$, and with

$$M(t) = M_0 + tM_1(t), \qquad M_0 := M(0),$$

we can reformulate (3.13) to

$$u'(t) = \frac{1}{t}M_0u(t) + F_1(t), \qquad 0 < t \leq 1, \tag{3.15}$$

where

$$F_1(t) := F(t) + M_1(t)u(t).$$

Since (3.15) is, formally, a system with constant coefficient matrix, its general solution is provided by (2.11). Let

$$\Phi_0(t) = \exp\left(M_0 \ln \frac{t}{\delta}\right) =: \left(\frac{t}{\delta}\right)^{M_0},$$

and $P_0 := P(0), Q_0 := Q(0) = I - P_0$; then Lemmas 2.2, 2.3, and 2.4 hold for Φ_0, P_0, Q_0, and the corresponding subspaces of M_0 and the general solution of (3.15) is

$$\begin{aligned} u(t) &= \Phi_0(t)\left\{c + \int_\delta^t \Phi_0^{-1}(s)F_1(s)ds\right\} \\ &= Q_0c + \Phi_0(t)P_0\left\{c + \int_\delta^t \Phi_0^{-1}(s)F_1(s)ds\right\} + Q_0\int_\delta^t F_1(s)ds. \end{aligned} \tag{3.16}$$

According to Lemma 3.1, it follows from $Q(\delta)u(\delta) = 0$ that $Q(t)u(t) = 0$ for all $t \in (0,1]$, which results in the following requirement on c

$$0 = Q(\delta)u(\delta) = Q(\delta)(Q_0c + P_0c) = Q(\delta)c.$$

Let us now associate with matrix M_0, projections P_1, P_2, R, H and the subspaces $N_0 := N(0)$, N_1, H_1, X_0, X_-, X_+, as defined in §2.2. Then the continuous solution $u \in C$ of (3.15) is

$$\begin{aligned} u(t) = {} & Q_0c + Q_0\int_\delta^t F_1(s)ds \\ & + Rc - \int_0^\delta R\Phi_0^{-1}(s)F_1(s)ds + t\int_0^1 P_1\Phi_0^{-1}(s)\Phi_0(1)F_1(st)ds \\ & + \Phi_0(t)\left\{P_2c + \int_\delta^t P_2\Phi_0^{-1}(s)F_1(s)ds\right\}, \qquad 0 < t \leq 1, \end{aligned} \tag{3.17}$$

where $P(\delta)c = c$, and, therefore,

$$\lim_{t\to 0} u(t) = Q_0P(\delta)c - Q_0\int_0^\delta F_1(s)ds + RP(\delta)c - \int_0^\delta R\Phi_0^{-1}(s)F(s)ds.$$

Moreover, the components $(Q_0 + R + P_2)P(\delta)c =: \gamma$ have to be chosen in a suitable way in order to satisfy boundary conditions (3.12b).

Since, $F_1(t)$ depends on $u(t)$, we interpret (3.17) as an operator equation, and use contraction techiques to show its solvability with respect to u. Therefore, let us discuss the integral operator $K : C[0,\delta] \to C[0,\delta]$,

$$\begin{aligned}(Ku)(t) :=& Q_0\gamma + Q_0 \int_\delta^t F_1(s)ds \\ &+R\gamma - \int_0^\delta R\Phi_0^{-1}(s)F_1(s)ds + t\int_0^1 P_1\Phi_0^{-1}(s)\Phi_0(1)F_1(st)ds \\ &+\Phi_0(t)\left\{P_2\gamma + \int_\delta^t P_2\Phi_0^{-1}(s)F_1(s)ds\right\}, \qquad 0 < t \le \delta,\end{aligned}$$

where $\gamma \in \mathbb{R}^m$ is a fixed vector. For u, $v \in C[0,\delta]$, we obtain

$$\begin{aligned}|(Ku)(t) - (Kv)(t)| \le& \left|\int_\delta^t Q_0M_1(s)(u(s)-v(s))ds\right| \\ &+\left|\int_0^\delta R\Phi_0^{-1}(s)M_1(s)(u(s)-v(s))ds\right| \\ &+\left|\int_0^1 tP_1\Phi_0^{-1}(s)\Phi_0(1)M_1(st)(u(st)-v(st))ds\right| \\ &+\left|\int_\delta^t \Phi_0(t)P_2\Phi_0^{-1}(s)M_1(s)(u(s)-v(s))ds\right|.\end{aligned} \tag{3.18}$$

We estimate (3.18) term by term; cf. [6]:

$$\begin{aligned}&\int_t^\delta |Q_0M_1(s)|ds \le \kappa_1\delta, \\ &\int_0^\delta |R\Phi_0^{-1}(s)M_1(s)|ds \le \kappa_2\delta, \\ t&\int_0^1 |P_1\Phi_0^{-1}(s)\Phi_0(1)M_1(st)|ds \le \kappa_3 t \le \kappa_3\delta, \\ &\int_t^\delta |\Phi_0(t)P_2\Phi_0^{-1}(s)M_1(s)|ds \le \kappa_4 \begin{cases} \delta\left(\dfrac{t}{\delta}\right)^{\sigma_{\min}}\left|\ln\dfrac{t}{\delta}\right|^n, & \sigma_{\min}\in(0,1), \\ t\left|\ln\frac{t}{\delta}\right|^{n+1}, & \sigma_{\min} = 1, \\ t\left|\ln\frac{t}{\delta}\right|^n, & \sigma_{\min} > 1, \end{cases} \\ &\le \kappa_4\delta,\end{aligned}$$

where $\sigma_{\min}$ is the smallest positive real part of the eigenvalues of M_0, and n is the dimension of the largest Jordan box associated with $\sigma_{\min}$. Consequently, it follows from (3.18), for sufficiently small $\delta > 0$, that

$$\|Ku - Kv\|_\infty \le \kappa\delta\,\|u - v\|_\infty < \|u - v\|_\infty$$

for any u, $v \in C[0,\delta]$, which means that K is contracting on $C[0,\delta]$ and has a unique fixed-point there. Equivalently, for any $\gamma = (Q_0 + R + P_2)P(\delta)c$, there exists a unique solution of (3.12a), $u(t) \in C[0,\delta]$, and this solution can be easily extended to $C[0,1]$.

We finally treat the question when u, given by (3.17), satisfies the boundary conditions (3.12b). The evaluation of (3.17) yields

$$\begin{aligned} u(0) &= (Q_0 + R)P(\delta)c \\ &\quad -Q_0 \int_0^\delta F_1(s)ds - \int_0^\delta R\Phi_0^{-1}(s)F(s)ds, \\ u(1) &= (Q_0 + R + \Phi_0(1)P_2)P(\delta)c \\ &\quad +Q_0 \int_\delta^1 F_1(s)ds - \int_0^\delta R\Phi_0^{-1}(s)F_1(s)ds \\ &\quad + \int_0^1 P_1\Phi_0^{-1}(s)\Phi_0(1)F_1(s)ds + \Phi_0(1)\int_\delta^1 P_2\Phi_0^{-1}(s)F_1(s)ds, \end{aligned}$$

and u is a unique solution of (3.12a), (3.12b), and (3.12c) if the linear system

$$Tc = \tilde{\beta}, \tag{3.19a}$$

$$T := \{B_0P_s(0)(Q_0 + R) + B_1P_s(1)(Q_0 + R + \Phi_0(1)P_2)\}P(\delta), \tag{3.19b}$$

$$\tilde{\beta} := \beta - B_1Q(1)G^{-1}(1)f(1) - B_0P_s(0)u_\beta(0) - B_1P_s(1)u_\beta(1), \tag{3.19c}$$

where

$$\begin{aligned} u_\beta(0) &:= -Q_0 \int_0^\delta F_1(s)ds - \int_0^\delta R\Phi_0^{-1}(s)F(s)ds, \\ u_\beta(1) &:= Q_0 \int_\delta^1 F_1(s)ds - \int_0^\delta R\Phi_0^{-1}(s)F_1(s)ds \\ &\quad + \int_0^1 P_1\Phi_0^{-1}(s)\Phi_0(1)F_1(s)ds + \Phi_0(1)\int_\delta^1 P_2\Phi_0^{-1}(s)F_1(s)ds, \end{aligned}$$

is uniquely solvable for the free components $(Q_0 + R)P(\delta)c$ and $P_2P(\delta)c$. Since

$$P_s(0)Q_0 = 0, \qquad \Phi_0(1)Q_0 = Q_0, \qquad \Phi_0(1)R = R,$$

we can rewrite (3.19b), and obtain

$$T = \{B_0P_s(0)R + B_1P_s(1)\Phi_0(1)(Q_0 + R + P_2)\}P(\delta). \tag{3.20}$$

We now recapitulate the results in the following theorem.

THEOREM 3.3. *Let $M(t)$ be differentiable at $t = 0$. Then the boundary value problem* (3.1) *has a unique solution $x \in C_\pi^1$, for any $f \in C$ and $\beta \in \mathbb{R}^k$, if and only if the matrix T, defined by* (3.19b), *satisfies the following conditions:*

$$\operatorname{rank}(T) = k = \dim(N_1) + \dim(X_+), \tag{3.21}$$

$$\ker(T) = \ker((Q_0 + R + P_2)P(\delta)). \tag{3.22}$$

Additionally, if the nullspace of $A(t)$ is constant, and we set $Q(t) \equiv Q_0$, then (3.20) *and* (3.22) *reduce to*

$$T = B_0P_s(0)R + B_1P_s(1)\Phi_0(1)(R + P_2)$$

and

$$\ker(T) = \ker(R + P_2) = N_0 \oplus H_1 \oplus X_-,$$

respectively.

COROLLARY 3.4. *Initial value problems are uniquely solvable if*

$$P_s(1)Q_0P(\delta) = 0, \qquad P_s(1)\Phi_0(1)P_2P(\delta) = 0.$$

This is the case when $P'(t) \equiv 0$ *and* $P_2 = 0$, *or equivalently, the matrix* $M(0) = -P(0)G^{-1}(0)B(0)$, *has no eigenvalues with positive real parts, and* $\det(\lambda A(0)+B(0)) = 0$ *and* $\lambda \neq 0$ *imply* $\mathrm{Re}(\lambda) < 0$.

Remarks.

1. If $A(t)$ and $B(t)$ are continuously differentiable at $t = 0$, so is $M(t)$.

2. Theorem 3.3 characterizes the existence of solutions $x \in C^1_\pi$. If $QG^{-1}f \in C^1_0$ and if all eigenvalues $\lambda = \sigma + i\eta$ with positive real parts, such that $\det(\lambda A(0)+B(0)) = 0$, satisfy the condition $\sigma > 1$, then $x \in C^1$.

REFERENCES

[1] E. GRIEPENTROG AND R. MÄRZ, *Differential-algebraic equations and their numerical treatment*, Teubner-Texte zur Mathematik, Band 88, Teubner Verlagsgesellschaft, Leipzig, Germany, 1986.

[2] F. R. DE HOOG AND R. WEISS, *Difference methods for boundary value problems with the singularity of the first kind,* SIAM J. Numer. Anal., 13 (1976), pp. 775–813.

[3] B. J. LEIMKUHLER, L. R. PETZOLD, AND C. W. GEAR, *Approximation methods for the consistent initialization of differential-algebraic equations*, SIAM J. Numer. Anal., 28 (1991), pp. 205–226.

[4] M. LENTINI AND R. MÄRZ, *The condition of boundary value problems in transferable differential-algebraic equations*, SIAM J. Numer. Anal., 27 (1990), pp. 1001–1015.

[5] E. B. WEINMÜLLER, *On the boundary value problem for systems of ordinary second order differential equations with a singularity of the first kind*, SIAM J. Math. Anal., 15 (1984), pp. 287–307.

[6] ———, *Stability of singular boundary value problems and their discretization by finite differences,* SIAM J. Numer. Anal., 26 (1989), pp. 180–213.

[7] R. MÄRZ, *On boundary value problems in differential-algebraic equations*, Appl. Math. Comp., 31 (1989), pp. 517–537.

[8] E. GRIEPENTROG AND R. MÄRZ, *Basic properties of some differential-algebraic equations,* Z. Anal. Anwendungen, 8 (1989), pp. 25–40.

[9] K. D. CLARK AND L. R. PETZOLD, *Numerical solution of boundary value problems in differential-algebraic equations*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 915–936.

SIAM J. MATH. ANAL.
Vol. 24, No. 1, pp. 216-222, January 1993

UNIFORM HARMONIC APPROXIMATION ON COMPACT SETS IN $\mathbb{R}^k$, $k \geqq 3$*

VLADIMIR ANDRIEVSKII†

Abstract. Estimates of the best uniform approximation of a function f given on some compact set $K \subset \mathbb{R}^k$, $k \geqq 3$, by harmonic polynomials are obtained. The results connect approximation properties of the function f with its structural properties and the geometry of K. This corresponds to well-known results for analytic functions.

Key words. approximation, harmonic function

AMS(MOS) subject classifications. 31A25, 30C85

1. Introduction. Let K be a compact set of the k-dimensional Euclidean space $\mathbb{R}^k$, $k \geqq 3$ with a connected complement $D := \mathbb{R}^k \backslash K$.

Denote by Har (K) the class of all functions continuous on K and harmonic at its interior points, and let H_n, $n = 0, 1, \ldots$, be the class of all harmonic polynomials of degree at most n.

For a function f given on K and $n = 0, 1, \ldots$, we set

$$\|f\|_K := \sup \{|f(\mathbf{x})|, \mathbf{x} \in K\},$$

$$E_{n,\Delta}(f, K) := \inf \{\|f - h\|_K, h \in H_n\}.$$

The main purpose of this paper is to establish estimates for the quantity $E_{n,\Delta}(f, K)$ depending—like the known results for analytic functions (the reader can find the whole survey in [4] and [5])—on the smoothness properties of the function f and the geometric structure of the compact set K. Similar problems for the case $k = 3$ were investigated in [8] and [1].

2. Main definitions and results. In all that follows $c, c_1, c_2, \ldots$ denote constants depending maybe on K, k, or other quantities inessential for our investigations. The same symbols may be used for different constants in the statements of different results.

For $\mathbf{x} = (x_1, \ldots, x_k) \in \mathbb{R}^k$, $\mathbf{y} = (y_1, \ldots, y_k) \in \mathbb{R}^k$, and $\delta > 0$, we set, as usual,

$$|\mathbf{x} - \mathbf{y}|^2 := \sum_{j=1}^{k} (x_j - y_j)^2,$$

$$B(\mathbf{x}, \delta) := \{\mathbf{y}: |\mathbf{x} - \mathbf{y}| < \delta\}, \qquad B(\delta) := B(0, \delta),$$

$$S(\mathbf{x}, \delta) := \{\mathbf{y}: |\mathbf{x} - \mathbf{y}| = \delta\}, \quad S(\delta) := S(0, \delta), \quad S := S(1),$$

$$d(\mathbf{x}, K) := \inf \{|\mathbf{x} - \mathbf{y}|, \mathbf{y} \in K\}.$$

The domain D is called a John domain (see, for example, [6]) if each point $\mathbf{x} \in D$ can be joined to ∞ with a Jordan curve $\gamma = \gamma(\mathbf{x})$ satisfying the following property. If γ is defined by $\mathbf{y} = \mathbf{y}(s)$, $s \in [0, \infty]$, $\mathbf{y}(0) = \mathbf{x}$, $\mathbf{y}(\infty) = \infty$, where s is the arc length, then for every $s > 0$ we require

$$d(\mathbf{y}(s), K) \geqq cs, \qquad c = \text{const} \in (0, 1]. \tag{2.1}$$

* Received by the editors May 28, 1991; accepted for publication (in revised form) May 5, 1992. This research was supported by the Alexander von Humboldt Foundation.

† Institute for Applied Mathematics and Mechanics of Ukrainian Academy of Sciences, ul. Rozy Luxemburg 74, Donetsk 340114, The Ukraine.

Let $\omega(\delta)$, $\delta > 0$ be a positive nondecreasing function, such that

$$\omega(+0) = 0; \qquad \omega(2\delta) \leqq c\delta, \quad \delta > 0.$$

We denote by $C^{\omega}(E)$ the class of functions f given on $E \subset \mathbb{R}^k$ for which

$$|f(\mathbf{x}) - f(\mathbf{y}) \leqq c\omega(|\mathbf{x}-\mathbf{y}|), \qquad \mathbf{x}, \mathbf{y} \in \mathrm{E}.$$

THEOREM 1. *Let K be a compact set whose complement D is a John domain. Then for $f \in C^{\omega}(K) \cap \operatorname{Har}(K)$ the following estimate holds*:

$$E_{n,\Delta}(f, K) \leqq c_1\omega(n^{-c}), \qquad n = 1, 2, \ldots, \tag{2.2}$$

where the constants c, $c_1 > 0$ are independent of n.

The proof of Theorem 1 is based on the following fact that is analogous to the procedure of "removal of the poles" suggested in the case of approximation of analytic functions by Keldysh (see, for example, [7, pp. 21-27]).

LEMMA 1. *Let $K \subset B(R)$ for some $R > 0$, and let $D = \mathbb{R}^k \backslash K$ be a John domain. For any $\mathbf{y} \in D \cap B(3R)$ and sufficiently small $\varepsilon \geqq 0$, there exists a function $q_{\mathbf{y}}(\mathbf{x}) = q_{\mathbf{y}}(\mathbf{x}, \varepsilon)$ harmonic in $\overline{B(3R)}$ with the following properties*:

$$||\mathbf{x}-\mathbf{y}|^{-k+2} - q_{\mathbf{y}}(\mathbf{x})| \leqq \varepsilon, \qquad \mathbf{x} \in K; \tag{2.3}$$

$$\|q_{\mathbf{y}}\|_{\overline{B(3R)}} \leqq \exp\left\{c_1 d^{-c_2} \ln\left[\frac{c_3}{d\varepsilon}\right]\right\}, \tag{2.4}$$

where $d := d(\mathbf{y}, K)$.

Moreover, we shall show how Lemma 1 can be applied to the proof of the harmonic analogue of the well-known Bernstein-Walsh theorem; see also Theorem 3.1 in [2].

THEOREM 2. *Let K be an arbitrary compact set with a simply connected complement, and let the function f be harmonic in some neighbourhood of K. Then for some $q \in (0, 1)$, the estimate*

$$E_{n,\Delta}(f, K) \leqq cq^n, \qquad n = 0, 1, \ldots, \tag{2.5}$$

holds, where the constant c is independent of n.

3. Proof of Lemma 1. To begin, we recall some facts that will be needed below (see, for example, [3, pp. 206, 213]).

Let the function F be given and continuous on the sphere $S(\mathbf{y}, R), \mathbf{y} \in \mathbb{R}^k$, $R > 0$. By the Laplace formula, the solution of Dirichlet's problem

$$\Delta Q(\mathbf{x}) = 0, \qquad \mathbf{x} \in \mathbb{R}^k \backslash S(\mathbf{y}, R),$$

$$Q(\mathbf{x}) = F(\mathbf{x}), \qquad \mathbf{x} \in S(\mathbf{y}, R)$$

can be represented by the series

$$Q(\mathbf{x}) = \sum_{j=0}^{\infty} \left(\frac{|\mathbf{x}-\mathbf{y}|}{\mathrm{R}}\right)^j Y_j(F, \xi), \qquad \mathbf{x} \in B(\mathbf{y}, R); \tag{3.1}$$

$$Q(\mathbf{x}) = \sum_{j=0}^{\infty} \left(\frac{R}{|\mathbf{x}-\mathbf{y}|}\right)^{j+k-2} Y_j(F, \xi), \qquad \mathbf{x} \in \mathbb{R}^k \backslash \overline{B(\mathbf{y}, R)}, \tag{3.2}$$

where $\xi := (\mathbf{x}-\mathbf{y})/|\mathbf{x}-\mathbf{y}|$, $Y_j(F, \xi)$, $j = 0, 1, \ldots,$ are the surface harmonics satisfying

$$\| Y_j(F, \cdot)\|_S \leqq c_1(j+1)^{k-2}\|F\|_{S(\mathbf{y},R)}. \tag{3.3}$$

Now let $\mathbf{y}$ and ε be given. Let $\gamma \subset D$ be an arc from the definition of John domain, i.e., an arc with endpoints $\mathbf{y}$ and ∞ satisfying (2.1). Without loss of generality, we may assume that $\gamma \cap [\mathbb{R}^k \backslash B(R)]$ is a part of some ray from the origin.

In the following c will be the constant from (2.1).

Setting $s_1 := d/4$, $s_\nu := (1+c/4)s_{\nu-1}$, $\nu = 2, 3, \ldots$, we introduce the sequence of points

$$\mathbf{y}_0 := \mathbf{y}, \quad \mathbf{y}_\nu := \mathbf{y}(s_\nu), \quad \nu = 1, 2, \ldots .$$

Our next aim is to construct a sequence of harmonic functions $q_0, q_1, \ldots$ with some special properties.

To begin with, put $q_0(\mathbf{x}) = |\mathbf{x} - \mathbf{y}_0|^{-k+2}$. Assuming that the function q_ν is given, we describe the procedure of construction of the next function $q_{\nu+1}$, $\nu = 0, 1, \ldots$.

Since the function q_ν is harmonic in $\mathbb{R}^k \backslash \{\mathbf{y}_\nu\}$ by (3.2), we have for $\mathbf{x} \in \mathbb{R}^k \backslash B(\mathbf{y}_{\nu+1}, cs_{\nu+1}/2)$, $\xi = (\mathbf{x} - \mathbf{y}_{\nu+1})/|\mathbf{x} - \mathbf{y}_{\nu+1}| \in S$,

$$q_\nu(\mathbf{x}) = \sum_{j=0}^{\infty} \left(\frac{cs_{\nu+1}}{2|\mathbf{x} - \mathbf{y}_{\nu+1}|} \right)^{j+k-2} Y_j(q_\nu, \xi).$$

Put

$$q_{\nu+1}(\mathbf{x}) = \sum_{j=0}^{n_{\nu+1}} \left(\frac{cs_{\nu+1}}{2|\mathbf{x} - \mathbf{y}_{\nu+1}|} \right)^{j+k-2} Y_j(q_\nu, \xi),$$

where the large enough number $n_{\nu+1}$ will be chosen later.

Introducing the notation $N_0 := d^{-k+2}$,

$$N_\nu := \|q_\nu\|_{S(\mathbf{y}_{\nu+1}, cs_{\nu+1}/2)}, \qquad \nu = 1, 2, \ldots ,$$

we obtain

$$N_\nu \leqq c_1 \sum_{j=0}^{n_\nu} (j+1)^{k-2} 2^j N_{\nu-1} \leqq c_2 n_\nu^{k-1} 2^{n_\nu} N_{\nu-1} \leqq 3^{n_\nu} N_{\nu-1}.$$

Moreover,

$$\|q_0 - q_1\|_K \leqq c_1\, d^{-k+2} \sum_{j=n_1+1}^{\infty} j^{k-2} 2^{-j} \leqq c_2\, d^{-k+2} n_1^{k-2} 2^{-n_1}, \tag{3.4}$$

and for $\nu \geqq 1$,

$$\begin{aligned} \|q_\nu - q_{\nu+1}\|_K &\leqq c_1 N_\nu \sum_{j=n_{\nu+1}+1}^{\infty} j^{k-2} 2^{-j} \\ &\leqq c_2 n_{\nu+1}^{k-2} 2^{-n_{\nu+1}} N_\nu. \end{aligned} \tag{3.5}$$

We consider the case $n_{\nu+1} := tn_\nu$, $\nu = 1, 2, \ldots$, for some integer $t > 1$.

We claim that there exist a sufficiently large $n_1 = n_1(\varepsilon)$ and t such that

$$\|q_\nu - q_{\nu+1}\|_K \leqq \left(\frac{\varepsilon}{2} \right)^{\nu+1}, \qquad \nu = 0, 1, \ldots . \tag{3.6}$$

Indeed, according to (3.4), inequality (3.6) is valid for $\nu = 0$ if we set

$$n_1 = c_1 \ln \frac{c_2}{\varepsilon d}, \tag{3.7}$$

where c_1 and c_2 are sufficiently large.

Further, since for $t>2$,

$$N_\nu \leqq 3^{n_\nu} N_{\nu-1} \leqq \cdots \leqq 3^{n_\nu+n_{\nu-1}+\cdots+n_1} N_0 \leqq 3^{n_\nu(1+t^{-1}+\cdots)} N_0 \leqq 3^{2n_\nu} N_0, \tag{3.8}$$

we have, applying (3.5) and (3.8) for $\nu \geqq 1$ and fixed $t \geqq 6$,

$$\|q_\nu - q_{\nu+1}\|_K \leqq c_1 t^{k-2} n_\nu^{k-2} 2^{-tn_\nu} 3^{2n_\nu} N_0 \leqq c_2 2^{-n_\nu} N_0 \leqq c_3 d^{-k-2} 2^{-n_1 t^{\nu-1}} \leqq \left(\frac{\varepsilon}{2}\right)^{\nu+1},$$

if the constants c_1 and c_2 from the relation (3.7) are sufficiently large.

Since our curve γ satisfies condition (2.1) there exists an integer $m = c_1 \ln(c_2/d)$ such that $B(3R) \cap B(\mathbf{y}_{m+1}, cs_{m+1}/2) = \phi$.

We claim that the function $q_\mathbf{y}(\mathbf{x}) := q_m(\mathbf{x})$ satisfies the desired conditions (2.3) and (2.4).

In fact, for $\mathbf{x} \in K$ and $\varepsilon \leqq 1$,

$$\left||\mathbf{x}-\mathbf{y}|^{-k+2} - q_\mathbf{y}(\mathbf{x})\right| \leqq |q_0(\mathbf{x}) - q_1(\mathbf{x})| + \cdots + |q_{m-1}(\mathbf{x}) - q_m(\mathbf{x})| \leqq \frac{\varepsilon}{2} + \left(\frac{\varepsilon}{2}\right)^2 + \cdots \leqq \varepsilon.$$

Apart from this, according to (3.8),

$$\|q_\mathbf{y}\|_{\overline{\mathrm{B(3R)}}} \leqq \mathrm{N_m} \leqq 3^{2\mathrm{n_m}} \mathrm{N_0} \leqq \exp\{c_1 t^m n_1 - c_2 \ln d\} \leqq \exp\left\{c_3\, d^{-c_4} \ln \frac{c_5}{\varepsilon d}\right\}.$$

Lemma 1 is proved.

With reasoning completely similar to the construction and study of the properties of the function q_1, we can obtain the following result.

For $E \subset \mathbb{R}^k$ and $\delta > 0$ set

$$E_\delta := \{\mathbf{x}\colon d(\mathbf{x}, E) < \delta\}.$$

LEMMA 2. *Let $D = \mathbb{R}^k \backslash K$ be a John domain. For any point $\mathbf{y} \in K_{3\delta} \cap D_\delta$, $\delta > 0$, there exists a function $\psi_{\delta,\mathbf{y}}(\mathbf{x})$ harmonic in $K_{3\delta}$, satisfying*

$$\|\psi_{\delta,\mathbf{y}}\|_{K_{3\delta}} \leqq c_1 \delta^{-k+2},$$

$$\left||\mathbf{x}-\mathbf{y}|^{-k+2} - \psi_{\delta,\mathbf{y}}(\mathbf{x})\right| \leqq c_2 \delta^3 |\mathbf{x}-\mathbf{y}|^{-k-1}, \qquad \mathbf{x} \in \mathbb{R}^k \backslash B(\mathbf{y}, c_3\delta).$$

4. Two auxiliary operators. We shall need the following two standard operators (for details, see [9], [10]).

For given $f \in \mathrm{Har}\,(K)$, denote by $\tilde{f} := \mathscr{E}_K f$ a function, defined on $\mathbb{R}^k$ such that

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}), \qquad \mathbf{x} \in K,$$

$$\tilde{f}(\mathbf{x}) = 0, \qquad \mathbf{x} \in \mathbb{R}^k \backslash K_{c_1},$$

$$\|\tilde{f}\|_{\mathbb{R}^k} \leqq c_2 \|f\|_K,$$

and in addition if $f \in C^\omega(K)$, then $\tilde{f} \in C^\omega(\mathbb{R}^k)$.

Furthermore, for a continuous function g given on $\mathbb{R}^k$ and $\delta > 0$ set

$$g_\delta(\mathbf{x}) = (U_\delta g)(\mathbf{x}) := \int_{\mathbb{R}^k} g(\mathbf{x} + \delta \mathbf{y}) \Phi(\mathbf{y})\, d\mathbf{y}, \qquad \mathbf{x} \in \mathbb{R}^k,$$

where

$$\Phi(\mathbf{y}) := \begin{cases} c \exp\{|\mathbf{y}|^2/(|\mathbf{y}|^2-1)\}, & 0 \leqq |\mathbf{y}| < 1; \\ 0 & |\mathbf{y}| > 1, \end{cases}$$

and the constant c is chosen such that $\int_{\mathbb{R}^k} \Phi(\mathbf{y})\, d\mathbf{y} = 1$.

The function g_δ has partial derivatives of all orders in $\mathbb{R}^k$, and if $g \in C^\omega(\mathbb{R}^k)$, then

$$\|g_\delta - g\|_{\mathbb{R}^k} \leqq c_1 \omega(\delta);$$

$$\|\Delta g_\delta\|_{\mathbb{R}^k} \leqq c_2 \omega(\delta)\delta^{-2}.$$

In addition, if g is harmonic on $\overline{B(\mathbf{x}, \delta)}$, then

$$g_\delta(\mathbf{x}) = g(\mathbf{x}).$$

5. Proof of Theorem 1. Let f be a given function, and let $\delta = \delta(n)$ be a sufficiently small number, the value of which will be determined below. Consider the function $\varphi = \varphi_\delta := U_\delta \mathscr{E}_K f$. According to Green's formula it can be written in the form

$$\varphi(\mathbf{x}) = -\frac{1}{(k-2)\sigma} \int_{D_\delta} \frac{\Delta\varphi(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|^{k-2}}\, d\mathbf{y}, \qquad \mathbf{x} \in \mathbb{R}^k,$$

where $\sigma = 2\pi^{k/2}/(\Gamma(k/2))$ is the total area of the unit sphere S.

It is easy to see that the function

$$u(\mathbf{x}) = u_\delta(\mathbf{x}) := \frac{1}{(k-2)\sigma} \int_{D_\delta \cap K_{3\delta}} \Delta\varphi(\mathbf{y}) \psi_{\delta,\mathbf{y}}(\mathbf{x})\, d\mathbf{y}$$
$$- \frac{1}{(k-2)\sigma} \int_{D_\delta \setminus K_{3\delta}} \frac{\Delta\varphi(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|^{k-2}}\, d\mathbf{y}, \qquad \mathbf{x} \in \mathbb{R}^k,$$

where $\psi_{\delta,\mathbf{y}}(\mathbf{x})$ is a function from Lemma 2, satisfies the conditions

$$u \in \operatorname{Har}(\overline{K_{2\delta}}), \quad \|u\|_{\overline{K_{2\delta}}} \leqq c_1 \|f\|_K, \quad \|u - f\|_K \leqq c_2 \omega(\delta).$$

Consider now the next function $v := U_\delta \mathscr{E}_{\overline{K_{2\delta}}} u$. From the properties of the operators U_δ and $\mathscr{E}_K$ we conclude that

$$\begin{aligned} &\|v - f\|_K \leqq c_1 \omega(\delta); \\ &\|\Delta v\|_{\mathbb{R}^k} \leqq c_2 \|f\|_K \delta^{-2}; \\ &\Delta v(\mathbf{x}) = 0, \qquad \mathbf{x} \in K_\delta; \\ &v(\mathbf{x}) = 0, \qquad \mathbf{x} \in \mathbb{R}^k \setminus B(3R) \end{aligned} \tag{5.1}$$

for some $R > 0$ such that $K \subset B(R)$.

Applying Green's formula to the function v we obtain

$$v(\mathbf{x}) = \frac{1}{(k-2)\sigma} \int_{B(3R)\setminus K_\delta} \frac{\Delta v(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|^{k-2}}\, d\mathbf{y}, \qquad \mathbf{x} \in \mathbb{R}^k.$$

Consider the function

$$q(\mathbf{x}) := \frac{1}{(k-2)\sigma} \int_{B(3R)\setminus K_\delta} \Delta v(\mathbf{y}) q_{\mathbf{y}} V(\mathbf{x})\, d\mathbf{y}, \qquad \mathbf{x} \in B(3R),$$

where $q_{\mathbf{y}}(\mathbf{x}) = q_{\mathbf{y}}(\mathbf{x}, \varepsilon)$, $\varepsilon = \delta^2\omega(\delta)$, is the function from Lemma 1. By virtue of Lemma 1, $q \in \operatorname{Har}(\overline{B(2R)})$ and

$$
\begin{aligned}
&\|v - q\|_K \leqq c_1\omega(\delta);\\
&\|q\|_{\overline{B(2R)}} \leqq c_2 \exp\{c_3\delta^{-c_4}\}.
\end{aligned}
\tag{5.2}
$$

By (3.1) and (3.3)

$$
q(\mathbf{x}) = \sum_{j=0}^{\infty} \left(\frac{|\mathbf{x}|}{2R}\right)^j Y_j(q, \xi); \quad \mathbf{x} \in B(2R), \quad \xi = \frac{\mathbf{x}}{|\mathbf{x}|};
$$

$$
\|Y_j(q, \cdot)\|_S \leqq c_5(j+1)^{k-2} \exp\{c_3\delta^{-c_4}\}.
$$

Considering the harmonic polynomial

$$
h_n(\mathbf{x}) := \sum_{j=0}^{n} \left(\frac{|\mathbf{x}|}{2R}\right)^j Y_j(q, \xi),
$$

we can estimate for $\mathbf{x} \in K$,

$$
|q(\mathbf{x}) - h_n(\mathbf{x})| \leqq c_6 \exp\{c_3\delta^{-c_4}\} \sum_{j=n+1}^{\infty} j2^{-j} \leqq c_7 \exp\{c_3\delta^{-c_4} - n \ln(\tfrac{3}{2})\}.
$$

Choosing $\delta := n^{-1/(2c_4)}$, we obtain by the inequalities (5.1) and (5.2) the desired estimate (2.2).

6. Proof of Theorem 2. The reasoning will be quite similar to that of Theorem 1. Without loss of generality, we may assume that $D = \mathbb{R}^k \backslash K$ is a John domain.

Fixing sufficiently small numbers $0 < d < \rho$ such that $f \in \operatorname{Har}(\overline{K_\rho})$, we consider the function $\varphi = \varphi_\delta := U_\delta \mathscr{E}_{\overline{K_\rho}} f$, where $0 < \delta < \rho - d$ will also be some fixed number.

By the properties of the operators U_δ and $\mathscr{E}_K$, we have

$$
\begin{aligned}
&\varphi(\mathbf{x}) = f(\mathbf{x}), \qquad \mathbf{x} \in \overline{K_d};\\
&\|\Delta\varphi\|_{\mathbb{R}^k} \leqq c_1;\\
&\varphi(\mathbf{x}) = 0, \qquad \mathbf{x} \in \mathbb{R}^k \backslash B(3R)
\end{aligned}
$$

for some R such that $K \subset B(R)$.

According to Green's formula we have

$$
\varphi(\mathbf{x}) = -\frac{1}{(k-2)\sigma} \int_{B(3R)\backslash K_d} \frac{\Delta\varphi(\mathbf{y})}{|\mathbf{x} - \mathbf{t}|^{k-2}}\, d\mathbf{y}, \qquad \mathbf{x} \in \mathbb{R}^k.
$$

Let n be large enough. Consider the function

$$
q(\mathbf{x}) := \frac{1}{(k-2)\sigma} \int_{B(3R)\backslash K_d} \Delta\varphi(\mathbf{y}) q_{\mathbf{y}}(\mathbf{x})\, d\mathbf{y}, \qquad \mathbf{x} \in \overline{B(3R)},
$$

where $q_{\mathbf{y}}(\mathbf{x}) = q_{\mathbf{y}}(\mathbf{x}, \varepsilon)$ is a function from Lemma 1, and $\varepsilon = \varepsilon(n)$ will be chosen below.

By the estimates (2.3) and (2.4) we obtain

$$
\begin{aligned}
&\|f - q\|_K \leqq c_1\varepsilon;\\
&\|q\|_{B(2R)} \leqq c_3\varepsilon^{-c}.
\end{aligned}
\tag{6.1}
$$

Since for $\mathbf{x} \in B(2R)$,

$$
q(\mathbf{x}) = \sum_{j=0}^{\infty} \left(\frac{|\mathbf{x}|}{2R}\right)^j Y_j(q, \xi), \qquad \xi = \frac{\mathbf{x}}{|\mathbf{x}|};
$$

$$
\|Y_j(q, \cdot)\|_S \leqq c_1(j+1)^{k-2}\varepsilon^{-c};
$$

we can obtain for the harmonic polynomial

$$h_n(\mathbf{x}) := \sum_{j=0}^{n} \left(\frac{|\mathbf{x}|}{2R}\right)^j Y_j(q, \xi),$$

the following estimate:

$$\|q - h_n\|_K \leqq c_1 \sum_{j=n+1}^{\infty} 2^{-j} j^{k-2} \varepsilon^{-c} \leqq c_2 2^{-n} n^{k-2} \varepsilon^{-c}. \tag{6.2}$$

It is easy to see that the desired estimate (2.5) follows from the inequalities (6.1) and (6.2) if $\varepsilon := q^n$, where $1 - q$ is positive and small enough.

Acknowledgment. This paper was completed while the author was visiting the Fachbereich Mathematik, Technische Universität Berlin, and the Mathematisches Institut, Justus-Liebig-Universität Giessen. The author wishes to thank the members of these universities for the pleasant mathematical atmosphere they offered him.

REFERENCES

[1] V. V. Andrievskii, *On the rate of harmonic approximation on compact sets in* $\mathbb{R}^3$, Ukrain. Mat. Zh., 41 (1989), pp. 1165-1169. (In Russian.)

[2] T. Bagby and N. Levenberg, *Bernstein theorems for harmonic functions*, Proc. of the Joint US-USSR Conference on Methods of Approximation Theory in Complex Analysis and Mathematical Physics, Leningrad, 1991, to appear.

[3] H. Behrens, P. L. Butzer, and A. Pavelke, *Limitierungsverfahren von Reihen mehrdimensionaler Kugelfunktionen und deren Saturationsverhalten*, Publ. RIMS, Kyoto Univ. Ser. A, 4 (1968), pp. 201-268.

[4] V. K. Dzyadyk, *Introduction to the Theory of Uniform Approximation of Functions by Polynomials*, Nauka, Moskow, 1977. (In Russian.)

[5] D. Gaier, *Lectures on Complex Approximation*, Birkhäuser, Boston, Basel, Stuttgart, 1987.

[6] O. Martio and J. Sarvas, *Injectivity theorems in plane and space*, Ann. Acad. Sci. Fenn. Ser. AI. Math., 4 (1978/1979), pp. 383-401.

[7] S. N. Mergelyan, *On completeness of a system of analytic functions*, Uspekhi Mat. Nauk, 8 (1953), pp. 3-63. (In Russian.)

[8] ———, *Harmonic approximation and approximate solution of the Cauchy problem for the Laplace equation*, Uspekhi Mat. Nauk, 11 (1956), pp. 3-26. (In Russian.)

[9] S. L. Sobolev, *Applications of Functional Analysis in Mathematical Physics*, American Mathematical Society, Providence, RI, 1963.

[10] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

SIAM J. MATH. ANAL.
Vol. 24, No. 1, pp. 223–233, January 1993

014

POSITIVE HERMITE INTERPOLATION BY QUADRATIC SPLINES*

AATOS LAHTINEN†

Abstract. Necessary and sufficient conditions are derived under which a quadratic spline preserves the positivity of a set of function values in the Hermite interpolation. As a corollary it is seen that positive interpolation is always possible over a set of nonnegative function values when a quadratic spline with suitable parameters is used.

Key words. quadratic spline, positive interpolation, parameters

AMS(MOS) subject classifications. 41A15, 41A29, 65D05, 65D07

1. Introduction. In the problem of positive interpolation a set of points where the measured or otherwise obtained ordinates are nonnegative is given, and the existence of a nonnegative interpolating function is considered. Such a function is needed in many practical situations, for instance in growth studies.

Spline functions offer good possibilities for the treatment of this problem. A simple linear spline as in [1] naturally gives a solution. However, it is usually desirable that the interpolating function has at least C^1-regularity. Under this condition the existence of positive interpolants has been studied for several types of splines (cf. [2], [4]–[6], [8], [9]). Especially [5] presents necessary and sufficient conditions under which the solution of the problem of positive interpolation is given by quadratic splines.

In some growth studies, the spline functions representing the boundary at different time levels must have a certain shape (cf. [3]). In these cases the derivatives at interpolating points also have to be taken into account. This leads to the problem of positive Hermite interpolation which is the subject of the present paper.

In §2 the problem is stated and a solution to the Hermite interpolation problem is given by using the quadratic splines of [7]. This solution has at most one additional breakpoint in each data interval. Section 3 considers positivity locally, that is, in a data interval. The situation is treated in terms of the endpoint derivatives m_i and m_{i+1} of the quadratic spline with a fixed additional breakpoint. Necessary and sufficient conditions to the local positivity are described as a region M_i in the $m_i m_{i+1}$-plane. The effect of the position of an additional breakpoint to the local positivity is considered in §4. Here the region M_i is taken as a function of the additional breakpoint. In §5 the main result, Theorem 1, gathers the local results together. Necessary and sufficient conditions are given for positive Hermite interpolation by quadratic splines in the whole interval. Finally, §6 establishes as a corollary that for nonnegative data a solution to the problem of positive interpolation by quadratic splines always exists.

2. The problem. We define at first the Hermite interpolation problem for quadratic splines as follows.

Let $[a, b]$ be an interval containing a mesh $(x_i)_1^n$ such that $a = x_1 < x_2 < \cdots < x_n = b$ and let $(y_i)_1^n$ and $(m_i)_1^n$ be real numbers. Find a quadratic spline s with the fewest number of breakpoints such that

$$s(x_i) = y_i, \qquad s'(x_i) = m_i, \quad i = 1, \ldots, n.$$

*Received by the editors June 10, 1991; accepted for publication (in revised form) May 5, 1992.

†Department of Mathematics, University of Helsinki, Hallituskatu 15, SF-00100 Helsinki, Finland.

Schumaker has in [7] given a solution to this problem. The construction is done separately to each subinterval $[x_i, x_{i+1}]$, $i = 1, \dots, n$.

If $m_i + m_{i+1} = 2\delta_i$, the solution s on $[x_i, x_{i+1}]$ is a parabola

$$s(x) = y_i + m_i(x - x_i) + \frac{m_{i+1} - m_i}{2\Delta x_i}(x - x_i)^2. \tag{2.1}$$

Here we have used the notation $\Delta x_i = x_{i+1} - x_i$, $\Delta y_i = y_{i+1} - y_i$, $\delta_i = \Delta y_i / \Delta x_i$.

If $m_i + m_{i+1} \neq 2\delta_i$, then s is on $[x_i, x_{i+1}]$ a quadratic spline with one breakpoint ξ_i, which can be freely chosen in the interval $]x_i, x_{i+1}[$.

$$s(x) = \begin{cases} y_i + m_i(x - x_i) + \frac{\mu_i - m_i}{2t_i \Delta x_i}(x - x_i)^2, & x_i \le x < \xi_i, \\ d_i + \mu_i(x - \xi_i) + \frac{m_{i+1} - \mu_i}{2(1 - t_i)\Delta x_i}(x - \xi_i)^2, & \xi_i \le x \le x_{i+1}, \end{cases} \tag{2.2}$$

where

$$t_i = \frac{\xi_i - x_i}{\Delta x_i}, \qquad \mu_i = 2\delta_i - t_i m_i - (1 - t_i) m_{i+1}, \tag{2.3}$$

$$d_i = (1 - t_i) y_i + t_i y_{i+1} + \tfrac{1}{2} t_i (1 - t_i) \Delta x_i (m_i - m_{i+1}). \tag{2.4}$$

Notice that $0 < t_i < 1, d_i = s(\xi_i)$ and $\mu_i = s'(\xi_i)$. The values d_i and μ_i are used in later considerations. They are a consequence of $s \in C^1[x_i, x_{i+1}]$ and of the interpolation conditions. For a fixed ξ_i, they are unique.

The quadratic spline s defined on an interval $[x_j, x_k] \subset [a, b]$ by (2.1) and (2.2) for $i = j, \dots, k - 1$ is called in the continuation *a solution of the Hermite interpolation problem*, or for short *a solution of the* HIP, on $[x_j, x_k]$ with data $(y_i)_j^k$ and $(m_i)_j^k$. This solution is characterized by having at most one breakpoint between interpolating points. The solution is, however, by no means unique because the additional breakpoints ξ_i remain as parameters expressed in terms of t_i.

Positive Hermite interpolation is a special case of Hermite interpolation. The formulation of this problem is as follows.

Let $[a, b]$ be an interval containing a mesh $(x_i)_1^n$ such that $a = x_1 < x_2 < \dots < x_n = b$, and let $(y_i)_1^n, y_i \ge 0, i = 1, \dots, n$ and $(m_i)_1^n$ be real numbers. Find a solution of the Hermite interpolation problem with data $(y_i)_1^n$ and $(m_i)_1^n$, which is nonnegative on $[a, b]$.

Our intention is to derive necessary and sufficient conditions for the existence of a positive Hermite interpolant and to examine the role of the breakpoints ξ_i in the process. As a corollary we will get results concerning positive interpolation.

Notice that the word "positive" is used as a synonym of nonnegative in this terminology.

3. Positivity on subintervals with fixed breakpoints. As a first step the positivity of the Hermite interpolation is considered on a subinterval $[x_i, x_{i+1}]$. Let there be given on $[x_i, x_{i+1}]$ values $y_i \ge 0, y_{i+1} \ge 0, m_i, m_{i+1}$ and a position $\xi_i \in]x_i, x_{i+1}[$ for an additional breakpoint if needed. The point ξ_i determines the value of parameter t_i via (2.3). Let then s be the solution of HIP on $[x_i, x_{i+1}]$ with this data so that the possible additional breakpoint is at ξ_i. The positivity of s is examined in different alternatives.

If $m_i + m_{i+1} = 2\delta_i$, the spline s is on $[x_i, x_{i+1}]$ a parabola defined in (2.1). Schmidt and Hess have shown in [5] that the positivity of such a spline s depends on the quantities

$$v_i = -\frac{2}{\Delta x_i}(y_i + \sqrt{y_i y_{i+1}}), \qquad i = 1, \ldots, n-1 \tag{3.1}$$

in the following way.

PROPOSITION 1. *If $m_i + m_{i+1} = 2\delta_i$, the solution s of the Hermite interpolation problem is positive on $[x_i, x_{i+1}]$ if and only if $m_i \geq v_i$.*

Thus we will concentrate for a while on the case where

$$(m_i, m_{i+1}) \notin L_i = \{(m_i, m_{i+1}) \mid m_i + m_{i+1} = 2\delta_i\}. \tag{3.2}$$

This means that the solution s of the HIP on $[x_i, x_{i+1}]$ is a quadratic spline with an additional breakpoint at a given point ξ_i as defined in (2.2).

A necessary condition for positivity is that

$$d_i = s(\xi_i) \geq 0. \tag{3.3}$$

This implies by (2.4) that positive Hermite interpolation with a fixed breakpoint is not possible for all data values. Therefore, we consider the relation of the derivative values m_i, m_{i+1} to other data values in a positive Hermite interpolant. In this setting there are given values $x_i, x_{i+1}, y_i, y_{i+1}, \xi_i$ and we will examine how these values determine a set M_i in the $m_i m_{i+1}$-plane so that every $(m_i, m_{i+1}) \in M_i$ gives a positive solution of HIP on $[x_i, x_{i+1}]$. The necessary condition for a pair (m_i, m_{i+1}) to produce a positive solution to the HIP is by (3.3),

$$(m_i, m_{i+1}) \in A_d = \{(m_i, m_{i+1}) \mid d_i \geq 0\}. \tag{3.4}$$

The positivity of s, a solution of the HIP with given data and the breakpoint at ξ_i is at first considered separately on the intervals $[x_i, \xi_i]$ and $[\xi_i, x_{i+1}]$, starting with the former. We have to treat several different cases. Notice that by the restriction (3.2) the points of the line L_i are excluded from these considerations.

Case 1. $(m_i, m_{i+1}) \in A_1 = \{(m_i, m_{i+1}) \in A_d \mid m_i = \mu_i\}$.

By (2.2) $s/[x_i, \xi_i]$ is of first degree and thus positive.

Case 2. $(m_i, m_{i+1}) \in A_2 = \{(m_i, m_{i+1}) \in A_d \mid m_i > \mu_i\}$.

When $m_i \neq \mu_i$, the restriction $s/[x_i, \xi_i]$ is a parabola given by (2.2). Its vertex is

$$(x^{(1)}, s(x^{(1)})) = \left(x_i - \frac{m_i}{\mu_i - m_i}\Delta x_i t_i,\ y_i - \frac{1}{2}\frac{m_i^2}{\mu_i - m_i}\Delta x_i t_i\right). \tag{3.5}$$

If $m_i > \mu_i$, then $s(x^{(1)}) \geq 0$ whence $s/[x_i, \xi_i] \geq 0$.

Case 3. $(m_i, m_{i+1}) \in A_3 = \{(m_i, m_{i+1}) \in A_d \mid m_i < \mu_i \ \wedge (m_i \geq 0 \vee \mu_i \leq 0)\}$.

In this case the vertex does not lie in the interval $]x_i, \xi_i[$. Therefore, $s/[x_i, \xi_i]$ is monotone and thus positive.

Case 4. $(m_i, m_{i+1}) \in A_d \setminus (A_1 \cup A_2 \cup A_3)$.

It is easy to see that $A_d \setminus (A_1 \cup A_2 \cup A_3) = \{(m_i, m_{i+1}) \in A_d \mid m_i < 0 \ \wedge \ \mu_i > 0\}$. This case is nonempty only when $y_i > 0$. In fact, if $y_i = 0$, then $s/[x_i, \xi_i]$ can be positive only if $m_i \geq 0$, which does not happen in Case 4.

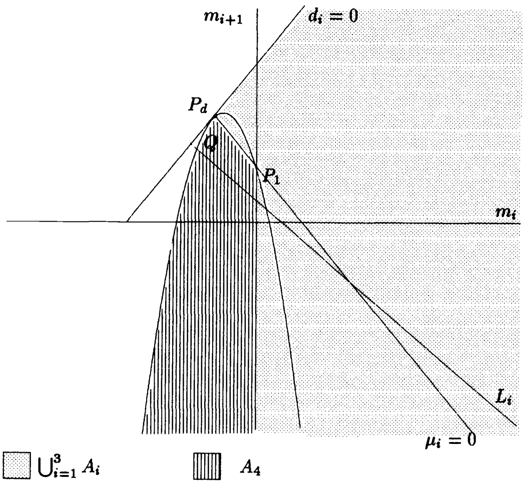


FIG. 1. *The sets* $\cup_{j=1}^3 A_j$ *and* A_4 *in a case where* $y_i y_{i+1} > 0$.

Now the minimum of the spline s is at the point $x^{(1)} \in \,]x_i, \xi_i[$. This means that $s/[x_i, \xi_i] \geq 0$ if and only if $s(x^{(1)}) \geq 0$. By (2.3) and (3.5) this condition is equivalent to

$$t_i \Delta x_i m_i^2 + 2y_i(1+t_i)m_i + 2y_i(1-t_i)m_{i+1} - 4y_i\delta_i \leq 0. \tag{3.6}$$

The closed region determined in the $m_i m_{i+1}$-plane by (3.6) is the inside of a parabola $m_{i+1} = p_i(m_i)$ opening downward. Here

$$p_i(m_i) = -\frac{t_i \Delta x_i}{2(1-t_i)y_i} m_i^2 - \frac{1+t_i}{1-t_i} m_i + \frac{2\delta_i}{1-t_i}. \tag{3.7}$$

The lines $d_i = 0$ and $m_i = \mu_i$ are both tangents to this parabola with points of tangency

$$P_d = (b_1, b_2) = \left(-\frac{2y_i}{t_i \Delta x_i}, \frac{2y_{i+1}}{(1-t_i)\Delta x_i}\right) \quad \text{and} \quad P_1 = \left(0, \frac{2\delta_i}{1-t_i}\right). \tag{3.8}$$

Furthermore, the parabola intersects the line $\mu_i = 0$ at the points P_d and P_1 and the line L_i at the points

$$Q = (v_i, w_i) \quad \text{and} \quad Q' = \left(-\frac{2}{\Delta x_i}(y_i - \sqrt{y_i y_{i+1}}), \frac{2}{\Delta x_i}(y_{i+1} - \sqrt{y_i y_{i+1}})\right), \tag{3.9}$$

where v_i is as (3.1) and

$$w_i = 2\delta_i - v_i = \frac{2}{\Delta x_i}(y_{i+1} + \sqrt{y_i y_{i+1}}). \tag{3.10}$$

Note that the points Q and Q' do not depend on t_i.

Thus $s/[x_i, \xi_i] \geq 0$ in Case 4 if and only if

$$(m_i, m_{i+1}) \in A_4 = \{(m_i, m_{i+1}) \in A_d \mid m_i < 0 \ \wedge \ \mu_i > 0 \ \wedge \ (3.6) \text{ is valid}\}. \tag{3.11}$$

If $y_i > 0$, the set A_4 is bounded by the parabola $m_{i+1} = p_i(m_i)$ up to the point P_d, by the line segment $P_d P_1$ and by the m_{i+1}-axis from the point P_1 downward (see Fig. 1). If $y_i = 0$, the set A_4 is empty.

The situation on the interval $[x_i, \xi_i]$ can be summed up as follows. If (3.2) is valid,

$$s/[x_i, \xi_i] \geq 0 \iff (m_i, m_{i+1}) \in \cup_{j=1}^{4} A_j \setminus L_i. \tag{3.12}$$

The part of the $m_i m_{i+1}$-plane that gives a positive solution to HIP on $[x_i, \xi_i]$ is the half plane $d_i \geq 0$ minus the line L_i and minus a wedge with its apex at P_d. The left side of the wedge is the line $d_i = 0$, and the right side a monotone part of the parabola $m_{i+1} = p_i(m_i)$ in case $y_i > 0$ (see Fig. 1). If $y_i = 0$, the right side of the wedge is the m_{i+1}-axis. The examination of the interval $[x_i, \xi_i]$ is now complete.

The situation on the interval $[\xi_i, x_{i+1}]$ can be treated quite in the same way as on $[x_i, \xi_i]$. Therefore, the discussion is shortened into two cases.

Case 5. $(m_i, m_{i+1}) \in A_5 = \{(m_i, m_{i+1}) \in A_d \mid m_{i+1} \leq 0 \ \vee \ \mu_i \geq 0\}$.

Analogously with Cases 1–3 it is seen that in Case 5 always $s/[\xi_i, x_{i+1}] \geq 0$.

Case 6. $(m_i, m_{i+1}) \in A_d \setminus A_5 = \{(m_i, m_{i+1}) \in A_d \mid m_{i+1} > 0 \ \wedge \ \mu_i < 0\}$.

This case is nonempty only when $y_{i+1} > 0$. In fact, if $y_{i+1} = 0$, then $s/[\xi_i, x_{i+1}]$ can be positive only if $m_{i+1} \leq 0$, which does not happen in Case 6.

In this case the minimum of the spline s is at a point $x^{(2)} \in]\xi_i, x_{i+1}[$. Thus $s/[\xi_i, x_{i+1}] \geq 0$ if and only if $s(x^{(2)}) \geq 0$. This is equivalent to

$$(1 - t_i)\Delta x_i m_{i+1}^2 - 2(2 - t_i)y_{i+1}m_{i+1} - 2t_i y_{i+1} m_i + 4y_{i+1}\delta_i \leq 0. \tag{3.13}$$

The closed region determined in the $m_i m_{i+1}$-plane by (3.13) is the inside of a parabola $m_i = q_i(m_{i+1})$ opening to the right. Here

$$q_i(m_{i+1}) = \frac{(1 - t_i)\Delta x_i}{2t_i y_{i+1}} m_{i+1}^2 - \frac{2 - t_i}{t_i} m_{i+1} + \frac{2\delta_i}{t_i}. \tag{3.14}$$

The lines $d_i = 0$ and $m_{i+1} = \mu_i$ are both tangents to this parabola with points of tangency P_d and, respectively,

$$P_2 = \left(\frac{2\delta_i}{t_i}, 0\right). \tag{3.15}$$

Furthermore, the parabola intersects the line $\mu_i = 0$ at the points P_d and P_2 and the line L_i at the same points Q and Q' as the parabola $m_{i+1} = p_i(m_i)$.

Thus $s/[\xi_i, x_{i+1}] \geq 0$ in Case 6 if and only if

$$(m_i, m_{i+1}) \in A_6 = \{(m_i, m_{i+1}) \in A_d \mid m_{i+1} > 0 \ \wedge \ \mu_i < 0 \ \wedge \ (3.13) \text{ is valid}\}. \tag{3.16}$$

If $y_{i+1} > 0$, the set A_6 is bounded by the parabola $m_i = q_i(m_{i+1})$ from the point P_d to the right, by the line segment $P_d P_2$ and by the m_i-axis from the point P_2 to the right. If $y_{i+1} = 0$, the set A_6 is empty.

The situation on the interval $[\xi_i, x_{i+1}]$ can be summed up as follows. If (3.2) is valid,

$$s/[\xi_i, x_{i+1}] \geq 0 \iff (m_i, m_{i+1}) \in \cup_{j=5}^{6} A_j \setminus L_i. \tag{3.17}$$

The part of the $m_i m_{i+1}$-plane that gives a positive solution to HIP on $[\xi_i, x_{i+1}]$ is the half plane $d_i \geq 0$ minus the line L_i and minus a wedge with its apex at P_d. The upper side of the wedge is the line $d_i = 0$ and the lower side a monotone part of the

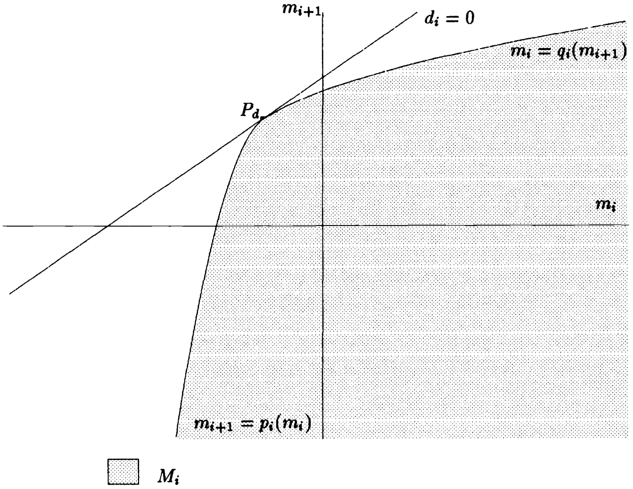


FIG. 2. *The set M_i in a case where $y_i y_{i+1} > 0$.*

parabola $m_i = q_i(m_{i+1})$ in case $y_{i+1} > 0$. If $y_{i+1} = 0$, the lower side of the wedge is the m_i-axis. The examination of the interval $[\xi_i, x_{i+1}]$ is complete.

The results of this paragraph can now be united. Denote

$$M_i = (\cup_{j=1}^{4} A_j) \cap (\cup_{j=5}^{6} A_j). \tag{3.18}$$

Note that the line L_i is no longer excluded.

PROPOSITION 2. *Let s be a solution of the Hermite interpolation problem on $[x_i, x_{i+1}]$ with data $y_i \geq 0, y_{i+1} \geq 0, m_i, m_{i+1}$, and let the possible additional breakpoint be at a point $\xi_i \in]x_i, x_{i+1}[$. Then $s \geq 0$ on $[x_i, x_{i+1}]$ if and only if*

$$(m_i, m_{i+1}) \in M_i.$$

Proof. If $(m_i, m_{i+1}) \notin L_i$, then the statement follows from (3.12), (3.17) and from the definition of M_i. If (m_i, m_{i+1}) lies in the line L_i, then by Proposition 1 $s \geq 0$ on $[x_i, x_{i+1}]$ if and only if $m_i \geq v_i$. On the other hand by the construction of M_i the set $\{(m_i, m_{i+1}) \in L_i \mid m_i \geq v_i\} \subset M_i$, the point $(v_i, 2\delta_i - v_i) = (v_i, w_i)$ being on the boundary of M_i. This proves the proposition. □

For later use we need a more detailed description of M_i (see Fig. 2). Firstly, suppose that $y_i y_{i+1} > 0$. Then we get from previous considerations that $(m_i, m_{i+1}) \in M_i$ if and only if

$$m_{i+1} \leq \begin{cases} p_i(m_i) & \text{if } m_i \leq b_1, \\ q_i^{-1}(m_i) & \text{if } m_i > b_1, \end{cases} \tag{3.19}$$

or equivalently

$$m_i \geq \begin{cases} p_i^{-1}(m_{i+1}) & \text{if } m_{i+1} \leq b_2, \\ q_i(m_{i+1}) & \text{if } m_{i+1} > b_2. \end{cases} \tag{3.20}$$

Here $(b_1, b_2) = P_d$ from (3.8), and p_i, q_i are determined by (3.7) and (3.14). Because p_i is monotone when $m_i \leq b_1$, it has an inverse function p_i^{-1} there. Similarly, q_i is monotone when $m_{i+1} \geq b_2$ and has an inverse function q_i^{-1}.

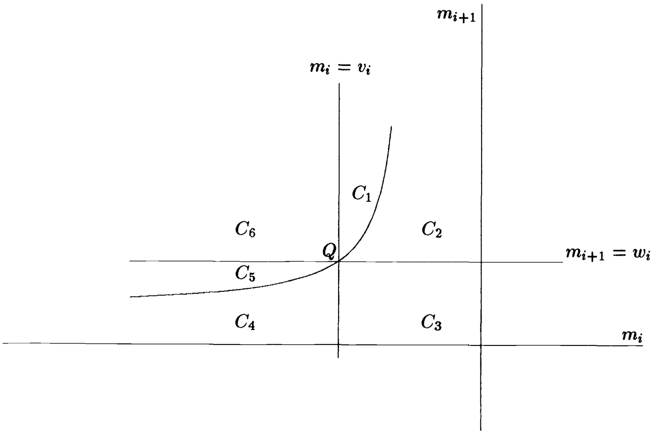


FIG. 3. *The partition of the $m_i m_{i+1}$-plane in the proof of Proposition* 3.

If $y_i = 0, y_{i+1} > 0$, then $(m_i, m_{i+1}) \in M_i$ if and only if (3.20) is valid with $p_i^{-1} = 0$. If $y_i > 0, y_{i+1} = 0$, then $(m_i, m_{i+1}) \in M_i$ if and only if (3.19) is valid with $q_i^{-1} = 0$. Finally, if $y_i = 0, y_{i+1} = 0$, then $(m_i, m_{i+1}) \in M_i$ if and only if

(3.21) $$m_i \geq 0 \ \wedge \ m_{i+1} \leq 0.$$

We notice that M_i is always closed, convex, and nonempty.

4. Positivity on subintervals. What we know up to now is exactly when the solution of the HIP is positive on $[x_i, x_{i+1}]$ if the place of the possible additional breakpoint ξ_i is fixed. In the problem of positive Hermite interpolation, however, the place of ξ_i is a free parameter. Therefore, we will next examine the dependence of the set $M_i = M_i(t_i)$ on the breakpoint ξ_i via t_i defined in (2.3).

PROPOSITION 3. *Let there be given on an interval $[x_i, x_{i+1}]$ data $y_i \geq 0, y_{i+1} \geq 0, m_i$ and m_{i+1}. There exists a positive solution of the Hermite interpolation problem on $[x_i, x_{i+1}]$ if and only if in the case where $y_i y_{i+1} > 0$,*

$$m_i > v_i \vee m_{i+1} < w_i \vee (m_i, m_{i+1}) = (v_i, w_i);$$

in the case where $y_i = 0, y_{i+1} > 0$,

$$m_i \geq 0;$$

in the case where $y_i > 0, y_{i+1} = 0$,

$$m_{i+1} \leq 0;$$

and in the case where $y_i = y_{i+1} = 0$,

$$m_i \geq 0 \ \wedge \ m_{i+1} \leq 0.$$

Proof. Firstly, suppose that $y_i y_{i+1} > 0$. We divide the $m_i m_{i+1}$-plane into four quadrants by the lines $m_i = v_i$ and $m_{i+1} = w_i$, which intersect at the point Q defined in (3.9). Furthermore, we divide the first and third quadrant by the branch of the hyperbola

(4.1) $$\Delta x_i m_i m_{i+1} - 2y_{i+1} m_i + 2y_i m_{i+1} = 0,$$

passing through the point Q. This branch is the track of the point P_d as a function of the parameter $t_i, 0 < t_i < 1$ (see Fig. 3).

We choose a point $P_o = (m_i^o, m_{i+1}^o)$ in each part of the $m_i m_{i+1}$-plane in a clockwise order and examine whether there exist values $t_i \in]0,1[$ so that $P_o \in M_i(t_i)$.

Case 1. $P_o \in C_1 = \{(m_i, m_{i+1}) \mid m_i > v_i \ \wedge \ m_{i+1} > (2y_{i+1}m_i)/(\Delta x_i m_i + 2y_i)\}$.

The boundary section $\partial M_i(t_i) \cap C_1$ is a monotone part of the parabola $m_{i+1} = p_i(m_i)$ (compare Figs. 1, 2, and 3). By (3.7) a point $P_o \in C_1$ belongs to $\partial M_i(t_i)$ if $t_i = t_i^{(1)}$, where

$$t_i^{(1)} = \frac{2y_i(2\delta_i - m_i^o - m_{i+1}^o)}{\Delta x_i m_i^{o^2} + 2y_i m_i^o - 2y_i m_{i+1}^o}, \tag{4.2}$$

supposing that $t_i^{(1)} \in]0,1[$. For the verification of this condition we first notice that on C_1 the numerator of $t_i^{(1)}$ is negative. For the denominator of $t_i^{(1)}$ we get a suitable upper bound,

$$\Delta x_i m_i^{o^2} + 2y_i m_i^o - 2y_i m_{i+1}^o < \Delta x_i v_i^2 + 2y_i v_i - 2y_i w_i = 0.$$

Thus $t_i^{(1)}$ is positive. Furthermore, it can be seen that on C_1,

$$\Delta x_i m_i^2 + 4y_i m_i - 4y_i \delta_i < \Delta x_i v_i^2 + 4y_i v_i - 4y_i \delta_i = 0,$$

and that the negativity of the left-hand side is equivalent to the inequality $t_i^{(1)} < 1$. It has been shown that $t_i^{(1)} \in]0,1[$.

The only boundary section of $M_i(t_i) \cap C_1$ that depends on t_i is the part of the parabola $m_{i+1} = p_i(m_i)$. From (3.7) we see that $p_i(m_i)$ is as a function of t_i monotonically increasing. This means that $t_i < u_i$ implies $M_i(t_i) \cap C_1 \subset M_i(u_i) \cap C_1$. So the point $P_o \in M_i(t_i) \cap C_1$ for every $t_i \in [t_i^{(1)}, 1[$. Thus a given point $P_o \in C_1$ lies in $M_i(t_i)$ if and only if the solution s of the HIP has an additional breakpoint ξ_i and

$$x_i + t_i^{(1)} \Delta x_i \le \xi_i < x_{i+1}. \tag{4.3}$$

Case 2. $P_o \in C_2 = \{(m_i, m_{i+1}) \mid m_{i+1} > w_i \ \wedge \ m_i \ge (2y_i m_{i+1})/(2y_{i+1} - \Delta x_i m_{i+1})\}$.

The boundary section $\partial M_i(t_i) \cap C_2$ is a monotone part of the parabola $m_i = q_i(m_{i+1})$ (compare Figs. 2 and 3). By (3.14) a point $P_o \in C_2$ belongs to $\partial M_i(t_i)$ if $t_i = t_i^{(2)}$, where

$$t_i^{(2)} = 1 + \frac{2y_{i+1}(2\delta_i - m_i^o - m_{i+1}^o)}{\Delta x_i m_{i+1}^{o^2} - 2y_{i+1} m_{i+1}^o + 2y_{i+1} m_i^o}. \tag{4.4}$$

As in Case 1 we see that $t_i^{(2)} \in]0,1[$. The only boundary section of $M_i(t_i) \cap C_2$ that depends on t_i is the part of the parabola $m_i = q_i(m_{i+1})$. As in Case 1 we see that $q_i(m_{i+1})$ is a monotone function of t_i, now decreasing. This means that $t_i < u_i$ implies $M_i(t_i) \cap C_2 \subset M_i(u_i) \cap C_2$. So the point $P_o \in M_i(t_i) \cap C_2$ for every $t_i \in [t_i^{(2)}, 1[$. Thus a given point $P_o \in C_2$ lies in $M_i(t_i)$ if and only if the solution s of the HIP has an additional breakpoint ξ_i and

$$x_i + t_i^{(2)} \Delta x_i \le \xi_i < x_{i+1}. \tag{4.5}$$

Case 3. $P_o \in C_3 = \{(m_i, m_{i+1}) \mid m_i \geq v_i \ \wedge \ m_{i+1} \leq w_i\}$.

The boundary $\partial M_i(t_i)$ is given by a strictly increasing concave function and the corner point of C_3, $Q = (v_i, w_i) \in \partial M_i(t_i)$ for every $t_i \in]0,1[$. This implies that $C_3 \subset M_i(t_i)$ for every $t_i \in]0,1[$. Thus a given point $P_o \in C_3$ lies in $M_i(t_i)$ for any additional breakpoint ξ_i,

$$x_i < \xi_i < x_{i+1}, \tag{4.6}$$

the solution s of the HIP has, and also if s has no additional breakpoint, i.e., if $m_i + m_{i+1} = 2\delta_i, m_i \geq v_i$.

Case 4. $P_o \in C_4 = \{(m_i, m_{i+1}) \mid m_i < v_i \ \wedge \ m_{i+1} \leq (2y_{i+1}m_i)/(\Delta x_i m_i + 2y_i)\}$.

As in Case 1 the boundary section $\partial M_i(t_i) \cap C_4$ is a monotone part of the parabola $m_{i+1} = p_i(m_i)$. Now $t_i > u_i$ implies $M_i(t_i) \cap C_4 \subset M_i(u_i) \cap C_4$. This means that the point $P_o \in M_i(t_i) \cap C_4$ for every $t_i \in]0, t_i^{(1)}]$. Thus a given point $P_o \in C_4$ lies in $M_i(t_i)$ if and only if the solution s of the HIP has an additional breakpoint ξ_i and

$$x_i < \xi_i \leq x_i + t_i^{(1)} \Delta x_i. \tag{4.7}$$

Case 5. $P_o \in C_5 = \{(m_i, m_{i+1}) \mid m_{i+1} < w_i \ \wedge \ m_i < (2y_i m_{i+1})/(2y_{i+1} - \Delta x_i m_{i+1})\}$.

As in Case 2 the boundary section $\partial M_i(t_i) \cap C_5$ is a monotone part of the parabola $m_i = q_i(m_{i+1})$. Now $t_i > u_i$ implies $M_i(t_i) \cap C_5 \subset M_i(u_i) \cap C_5$. This means that the point $P_o \in M_i(t_i) \cap C_5$ for every $t_i \in]0, t_i^{(2)}]$. Thus a given point $P_o \in C_5$ lies in $M_i(t_i)$ if and only if the solution s of the HIP has an additional breakpoint ξ_i and

$$x_i < \xi_i \leq x_i + t_i^{(2)} \Delta x_i. \tag{4.8}$$

Case 6. $P_o \in C_6 = \mathbf{R}^2 \setminus \cup_{j=1}^5 C_j$.

It is easy to see that in the case, where $y_i y_{i+1} > 0$,

$$C_6 = \{(m_i, m_{i+1}) \mid m_i \leq v_i \ \wedge \ m_{i+1} \geq w_i \ \wedge \ (m_i, m_{i+1}) \neq (v_i, w_i)\}.$$

By the arguments given in Case 3 it must be $M_i(t_i) \cap C_6 = \emptyset$ for every $t_i \in]0,1[$. Thus no $P_o \in C_6$ can give a positive Hermite interpolant.

The case $y_i y_{i+1} > 0$ is now completely examined. It has been shown that for any given data the solution of the HIP is positive in Cases 1, 2, 4, and 5 with a suitable choice of additional breakpoint and in Case 3 independent of the breakpoint. This proves the proposition in this case.

The treatment is valid with obvious modifications in cases where $y_i = 0$ or $y_{i+1} = 0$. The proposition is proved. □

From the proof we get also exact conditions for the choice of additional breakpoints. We state separately the case where the positivity is independent of the choice of breakpoint.

PROPOSITION 4. *Let there be given on an interval* $[x_i, x_{i+1}]$ *data* $y_i \geq 0, y_{i+1} \geq 0, m_i$ *and* m_{i+1}. *All solutions of the Hermite interpolation problem are positive on* $[x_i, x_{i+1}]$ *if and only if*

$$m_i \geq v_i \ \wedge \ m_{i+1} \leq w_i.$$

5. Positive Hermite interpolation. Now it can be stated when a quadratic spline defined by (2.1) and (2.2) will produce positive Hermite interpolation either independent of the choice of additional breakpoints or with a suitable choice of them.

THEOREM 1. *Let there be given on an interval $[a,b]$ a mesh $(x_i)_1^n$, $a = x_1 < x_2 < \cdots < x_n = b$ and real numbers $(y_i)_1^n, y_i \geq 0, i = 1, \ldots, n$ and $(m_i)_1^n$. There exists a quadratic spline as in (2.1) and (2.2) giving a positive solution of the Hermite interpolation problem on $[a,b]$ with this data if and only if at every point x_i, where $y_i = 0$,*

$$m_1 \geq 0 \; if \; i = 1, \quad m_i = 0 \; if \; i = 2, \ldots, n-1, \quad m_n \leq 0 \; if \; i = n,$$

and in every interval $[x_i, x_{i+1}]$, where $y_i y_{i+1} > 0$,

$$m_i > v_i \;\vee\; m_{i+1} < w_i \;\vee\; (m_i, m_{i+1}) = (v_i, w_i).$$

This result follows directly from Proposition 3. The conditions of Theorem 1 can also be verified by examining the validity of the following restrictions:

For $i = 1$ $\quad m_1 \geq 0$ $\quad$ if $y_1 = 0$;
For $i = 2, \ldots, n-1$ $\quad m_i = 0$ $\quad$ if $y_i = 0$ else
$m_i \leq w_{i-1}$ $\quad$ if $m_{i-1} = v_{i-1}$ else
$m_i < w_{i-1}$ $\quad$ if $m_{i-1} < v_{i-1}$;
For $i = n$ $\quad m_n \leq w_{n-1}$ $\quad$ if $m_{n-1} = v_{n-1} \;\vee\; y_n = 0$ else
$m_n < w_{n-1}$ $\quad$ if $m_{n-1} < v_{n-1}$.

Similarly we get the following result from Proposition 4.

THEOREM 2. *Let there be given on an interval $[a,b]$, a mesh $(x_i)_1^n$, $a = x_1 < x_2 < \cdots < x_n = b$, and real numbers $(y_i)_1^n, y_i \geq 0, i = 1, \ldots, n$ and $(m_i)_1^n$. Every quadratic spline as in (2.1) and (2.2), which is a solution of the Hermite interpolation problem with this data, is nonnegative on $[a,b]$ if and only if*

$$v_1 \leq m_1, \quad v_i \leq m_i \leq w_{i-1}, \quad i = 2, \ldots, n-1, \quad m_n \leq w_{n-1}.$$

In a situation where the conditions of Theorem 1 are valid but the conditions of Theorem 2 are not, some but not all additional breakpoints produce a positive Hermite interpolant. The proof of Proposition 3 gives necessary and sufficient conditions (formulas (4.2)–(4.8)) for the choice of the additional breakpoints.

The results imply that the problem of positive Hermite interpolation is not always solvable by using quadratic splines as in (2.1) and (2.2). In fact, if at an inner point x_i where $y_i = 0$ we have $m_i \neq 0$, then no Hermite interpolant of any kind can be positive in the neighbourhood of x_i.

On the other hand, by allowing in the quadratic spline more than one breakpoint between interpolating points we can in all other cases obtain a solution to the problem of positive Hermite interpolation. It is sufficient to consider the case where on an interval $[x_i, x_{i+1}]$ we have $y_i y_{i+1} > 0$, but the conditions of Theorem 1 are not fulfilled. In this situation we add a new interpolating point z_i to the interval $]x_i, x_{i+1}[$ and set both the function value and the derivative value to be zero at z_i. Now Theorem 1 shows that for this augmented data set we get a positive Hermite interpolant, which of course is also a solution to the original problem. This gives us the following modification of Theorem 1.

COROLLARY 1. *Let there be given on an interval $[a,b]$ a mesh $(x_i)_1^n$, $a = x_1 < x_2 < \cdots < x_n = b$ and real numbers $(y_i)_1^n, y_i \geq 0, i = 1, \ldots, n$ and $(m_i)_1^n$. There exists a nonnegative quadratic spline s on $[a,b]$ with $s(x_i) = y_i$ and $s'(x_i) = m_i$, $i = 1, \ldots, n$ if and only if at every point x_i where $y_i = 0$,*

$$m_1 \geq 0 \quad if\, i = 1; \quad m_i = 0 \quad if\, i = 2, \ldots, n-1; \quad m_n \leq 0 \quad if\, i = n.$$

6. Positive interpolation. Finally we would like to remark that because the set M_i is never empty, we can see from the theorems that the quadratic splines we use always give a solution to the problem of positive interpolation.

COROLLARY 2. *Let there be given on an interval* $[a,b]$ *a mesh* $(x_i)_1^n$, $a = x_1 < x_2 < \cdots < x_n = b$ *and real numbers* $(y_i)_1^n, y_i \geq 0, i = 1, \ldots, n$. *There exists a quadratic spline as in* (2.1) *and* (2.2) *that is a positive interpolant to this data.*

In fact, by choosing $m_i = 0, i = 1, \ldots, n$, for instance, we get for every set of additional breakpoints a positive Hermite interpolant, which is of course also a positive interpolant.

REFERENCES

[1] C. DE BOOR, *A Practical Guide to Splines*, Springer-Verlag, New York, Berlin, 1978, p. 392.

[2] C. DE BOOR AND J. DANIEL, *Splines with non-negative B-spline coefficients*, Math. Comput., 28 (1974), pp. 565–568.

[3] A. LAHTINEN, *On the taper curves in the forest industry*, in Proceedings of the Fourth European Conference on Mathematics in Industry, Hj. Wacker and W. Zulehner, eds., B. G. Teubner, Stuttgart, and Kluwer Academic Publishers, Dordrecht, 1991, pp. 323–327

[4] J. SCHMIDT, *Results and problems in shape preserving interpolation and approximation with polynomial splines*, in Splines in Numerical Analysis, J. Schmidt and H. Späth, eds., Akademie Verlag, Berlin, 1989, pp. 159–170.

[5] J. SCHMIDT AND W. HESS, *Positive interpolation with rational quadratic splines*, Computing, 38 (1987), pp. 261–267.

[6] ———, *Positivity of cubic polynomials on intervals and positive spline interpolation*, BIT, 28 (1988), pp. 340–352.

[7] L. SCHUMAKER, *On shape preserving quadratic spline interpolation*, SIAM J. Numer. Anal., 20 (1983), pp. 854–864.

[8] H. SPÄTH, *Spline-Algorithmen zur Konstruktion glatter Kurven und Flächen*, 3. Aufl., Oldenburg Verlag, München, Wien, 1983, p. 134

[9] U. WEVER, *Non-negative exponential splines*, Comput. Aided Design, 20 (1988), pp. 11–16.

SIAM J. MATH. ANAL.
Vol. 24, No. 1, pp. 234–245, January 1993

015

SUBDIVISION SCHEMES DETERMINED BY COEFFICIENTS OF A HURWITZ POLYNOMIAL*

I. YAD-SHALOM†

Abstract. The functional equation

$$E(t) = \sum_{i=0}^{k} a_i E(2t-i), \qquad \sum_{i=-\infty}^{\infty} E(t-i) = 1$$

is known to have a unique compactly supported continuous solution, given that $\{a_i\}_{i=0}^{k}$ is a positive sequence satisfying $\Sigma_i a_{2i} = \Sigma_i a_{2i-1} = 1$. This solution, $E(t)$, is also obtained by a subdivision algorithm. This paper discusses the particular case where $\Sigma_{i=0}^{k} a_i z^i$ is a Hurwitz polynomial, i.e., all zeros are in the left half plane.

The problem of Lagrange interpolation from the space $\{E(\cdot - i)\}$, $i \in Z$, was recently analyzed by Goodman and Micchelli [*SIAM J. Math. Anal.*, 23 (1992), pp. 766–784]. In particular, they showed that the problem is solvable under a condition similar to the B-spline case.

Here an alternative analysis is introduced, and results on the problem of Hermite interpolation are stated. The local linear independence of $\{E(\cdot - i)\}$, $i \in Z$, is discussed also for the case where $\Sigma_{i=0}^{k} a_i z^i$ is not a Hurwitz polynomial.

Key words. stationary subdivision schemes, functional equations, Hurwitz polynomials, total nonnegativity, Hermite interpolation

AMS(MOS) subject classifications. primary 41A, 39B; secondary 15

1. Introduction.

The functional equation

$$E(t) = \sum_{i=0}^{k} a_i E(2t-i)$$

has found important applications in the areas of computer aided geometric design and multiresolution analysis. The uniform normalized B-spline of order k (degree $k-1$) with integer knots $\{0, \ldots, k\}$ satisfies the equation above with the mask

$$a_i = 2^{-(k-1)} \binom{k}{i}, \qquad 0 \le i \le k\ .$$

Generally, solutions of such equations (if they exist) do not have an explicit form and they are approximated by subdivision schemes. These schemes originate in a well-known algorithm due to Lane and Riesenfeld. Subdivision schemes are useful for curve (surface) generation, and for a detailed discussion of this subject see, e.g., [DGL], [D], [MP], [CDM], and [DL].

Wavelets are deeply related to the equation above, and we refer the reader to [DAUB] for a discussion of orthonormal wavelets, and to [CW] and [M] for a discussion of prewavelets.

*Received by the editors July 29, 1991; accepted for publication (in revised form) May 1, 1992.

†Raymond and Beverly Sackler Faculty of Exact Sciences, School of Mathematical Sciences, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel.

Here we restrict ourselves to a certain class of equations (containing the B-spline equations), and we mainly analyze determinantal and sign properties of their compactly supported continuous solution. As we clarify now, some properties of the solution, $E(t)$, are particularly significant for the research areas mentioned above.

(i) [GM]. $\{E(t-i)\}_{i=-\infty}^{\infty}$ is a totally nonnegative sequence. Such a basis is particularly shape-preserving, and for a discussion of such representations we refer the reader to [G].

(ii) [GM]. $\{E(t-i)\}_{i=-\infty}^{\infty}$ is a locally linearly independent sequence in the sense of Definition 3.1. Such sequences are known to be L^2-stable, and this stability is important in multiresolution analysis.

Our motivation for this paper arises from an analysis done by [GM] and in the following we briefly survey their results.

The polynomial

$$A(z)=\sum_{i=0}^{k} a_i z^i, \qquad \{a_i\}_{i=0}^{k} \subset R \tag{1.1}$$

is termed a Hurwitz polynomial if all its zeros satisfy Re $z<0$ (see [GANT]). We will also call the associated sequence $\{a_i\}_{i=0}^{k}$, a Hurwitz sequence. In the particular case, where the zeros are real and negative, $\{a_i\}_{i=0}^{k}$ is termed a Polya frequency sequence.

Imposing the condition

$$A(1)=2, \qquad A(-1)=0, \tag{1.2}$$

we find that the Hurwitz sequence $\{a_i\}_{i=0}^{k}$ is positive ($a_i>0,\ 0\le i\le k$) and the functional equation

$$E(t)=\sum_{i=0}^{k} a_i E(2t-i)\ , \qquad \sum_{i=-\infty}^{\infty} E(t-i)=1 \tag{1.3}$$

has a unique continuous compactly supported solution (see also [MP], [CDM]). The function $E(t)$ is vanishing outside of $(0,k)$ and satisfies a variation diminishing property

$$S^-\left(\sum_{i=-\infty}^{\infty} \alpha_i E(t-i)\right) \le S^-\{\alpha_i\}, \qquad t\in(-\infty\ ,\ \infty), \quad \{\alpha_i\}_{i=\infty}^{\infty}\subset R, \tag{1.4}$$

where S^- counts strong sign changes of functions, sequences, respectively. Given a sequence $\{x_{\ell_\nu}\}_{\nu=0}^{r}$, $x_{\ell_\nu}<x_{\ell_{\nu+1}}$, where $\{\ell_\nu\}$ is a strictly increasing sequence of integers, the matrix

$$M\begin{pmatrix} x_{\ell_0} & \cdots & x_{\ell_r} \\ E(\cdot-\ell_0) & \cdots & E(\cdot-\ell_r)\end{pmatrix}_{i,j} = E(x_{\ell_i}-\ell_j) \tag{1.5}$$

is totally nonnegative, i.e., all its minors are nonnegative. We denote the associated determinant by

$$D\begin{pmatrix} x_{\ell_0} & \cdots & x_{\ell_r} \\ E(\cdot-\ell_0) & \cdots & E(\cdot-\ell_r)\end{pmatrix},$$

and, in particular,

$$(1.6)\qquad D\begin{pmatrix} x_{\ell_0} & \cdots & x_{\ell_r} \\ E(\cdot-\ell_0) & \cdots & E(\cdot-\ell_r) \end{pmatrix} \geq 0, \qquad x_{\ell_\nu} < x_{\ell_{\nu+1}}\,.$$

Moreover, (1.6) is positive if and only if

$$(1.7)\qquad x_{\ell_\nu} \in \operatorname{supp}(E(\cdot-\ell_\nu)) = (\ell_\nu\ ,\ \ell_\nu + k)$$

which is a well-known result for B-splines (see, e.g., [SCH]). Here and throughout this paper the support of a function is defined by $\operatorname{supp}(f) := \{x \mid f(x) \neq 0\}$.

The analysis in this paper is not based on [GM], and we believe that our proofs are simpler. First we establish in §3 the local linear independence of $\{E(\cdot - i)\}$, given that $\{a_i\}_{i=0}^k$ is a Hurwitz sequence. We also analyze the local linear independence where $\{a_i\}_{i=0}^k$ is a positive sequence and where $\{a_i\}_{i=0}^k$ is a mask of an interpolatory subdivision scheme. In §4 we introduce an alternative analysis for the problem of Lagrange interpolation, where $\{a_i\}_{i=0}^k$ is a Hurwitz sequence, and in §5 we discuss the Hermite interpolation problem.

Our proofs are mainly based on arguments of subdivision and arguments of Weak Tschebycheff (WT-) systems. A survey of WT-systems is found in [SCH] §2, and we completely follow the approach there. The results of subdivision theory are mainly taken from [D], which is a review paper summarizing known results (see also, e.g., [CDM], [DL], [DGL], [MP]). Other key results are Orlando's formula [GANT] and the total positivity of a Hurwitz matrix [KEMP], [AS], which are explained in §2.

We comment that subdivision schemes determined by a Hurwitz polynomial are a subclass of those that are analyzed in [Y]. The properties investigated there are Lipschitz continuity, variation diminishing properties, and bell-shaped refinable functions.

2. Subdivision preliminaries. In the following we clarify some well-known results related to subdivision schemes and functional equations. Unless otherwise stated, the results are taken from [D] (and the contributors are credited there).

A subdivision scheme is given by the formula

$$(2.1)\qquad f_i^{n+1} = \sum_{j=-\infty}^{\infty} a_{i-2j} f_j^n,$$

where $\{f_i^n\}_{i=-\infty}^{\infty} \subset R^d$ denotes the control points at level n, and $\{a_i\}_{i=0}^k$ ($a_0 \neq 0$, $a_k \neq 0$, $a_i = 0$ for $i < 0$, $i > k$) is the mask of the scheme.

Definition 2.1 (Convergence). Let (2.1) apply recursively to the initial data $\{f_i^0 = \delta_{i,0}\}_{i=-\infty}^{\infty} \subset R$. The scheme is said to converge if there exists a nontrivial continuous function $E(t)$ such that $|E(2^{-n}i) - f_i^n|$ tends to zero uniformly on i and on n.

The function $E(t)$ is supported on $(0, k)$ and is termed the refinable function of the scheme. The following fundamental theorem summarizes the relation between convergent schemes and functional equations.

Theorem 2.2. *Let $E(t)$ be the refinable function of a convergent scheme determined by $\{a_i\}_{i=0}^k$. Then $E(t)$ is the unique continuous, compactly supported solution to*

$$(2.2)\qquad E(t) = \sum_{i=0}^{k} a_i E(2t - i),$$

which satisfies also

$$\sum_{i=-\infty}^{\infty} E(t-i) = 1. \tag{2.3}$$

The sum rule

$$\sum_i a_{2i} = \sum_i a_{2i+1} = 1 \tag{2.4}$$

is a necessary condition for a convergent scheme, and it is equivalent to saying that the characteristic polynomial $A(z)$ given by

$$A(z) = \sum_{i=0}^{k} a_i z^i \tag{2.5}$$

satisfies

$$A(1) = 2, \qquad A(-1) = 0. \tag{2.6}$$

The following theorem is central for analyzing the smoothness of $E(t)$.

THEOREM 2.3. *Assume $k > 1$. Let $\{a_i^{(1)}\}_{i=0}^{k-1}$ be given by*

$$A^{(1)}(z) = 2(1+z)^{-1}A(z) = \sum_{i=0}^{k-1} a_i^{(1)} z^i \ . \tag{2.7}$$

If $\{a_i^{(1)}\}_{i=0}^{k-1}$ determines a convergent scheme, then $\{a_i\}_{i=0}^{k}$ also does. Moreover, let $E(t)$, $F(t)$ be the refinable functions associated with $\{a_i\}_{i=0}^{k}$, $\{a_i^{(1)}\}_{i=0}^{k-1}$; correspondingly then $E(t) \in C^1$ and

$$E'(t) = F(t) - F(t-1) \ . \tag{2.8}$$

The scheme determined by $\{a_i^{(1)}\}_{i=0}^{k-1}$ is termed the divided difference scheme. More generally, if $k > m$ and if $A(z)$ has an m-fold zero at $z = -1$, then the coefficients of $2^m \ (z+1)^{-m} \ A(z)$ determine the divided difference scheme of order m. It is also true that if the divided difference scheme of order m converges, then $E(t) \in C^m$. The mask of the divided difference scheme of order m, $\{a_i^{(m)}\}$ is given explicitly by

$$\sum_{i=0}^{k-m} a_i^{(m)} z^i = 2^m (z+1)^{-m} \left(\sum_{i=0}^{k} a_i z^i \right) . \tag{2.9}$$

Let the matrices A_0 and A_1 be defined by

$$\begin{aligned} (A_0)_{i,j} &= a_{i-2j+k-1}, \qquad 0 \le i,j \le k-1, \\ (A_1)_{i,j} &= a_{i-2j+k}, \qquad 0 \le i,j \le k-1, \end{aligned} \tag{2.10}$$

and let $\overline{f_i^n}$ denote the vector $(f_{i-(k-1)}^n, \dots, f_i^n)$. Then the following is true.

THEOREM 2.4.

(i) *The values at level* n, $\{f_i^n\}$, *satisfy*

$$\sum_i f_i^0 E(t-i) = \sum_i f_i^n E(2^n t - i), \tag{2.11}$$

and $\overline{f_j^n}$ *determines the behavior of* (2.11) *on* $[j2^{-n}, (j+1)2^{-n}]$.

(ii) *The transformation* (2.1) *is given also by* $\overline{f_{2i}^{n+1}} = A_0\overline{f_i^n}$, $\overline{f_{2i+1}^n} = A_1\overline{f_i^n}$. *Hence, for each* $i \in Z$ *there exist* $j \in Z$ *and a sequence* $\{i_\ell\}_{\ell=0}^n$ *such that*

$$\overline{f_i^n} = \prod_{\ell=1}^n A_{i_\ell}\overline{f_j^0}, \qquad i_\ell \in \{0,1\}. \tag{2.12}$$

(iii) [DGL]. A_0 *and* A_1 *have* $k-1$ *common eigenvalues counting multiplicities. The kth eigenvalue of* A_0 *is* a_0, *and the kth eigenvalue of* A_1 *is* a_k. (*Note that since* $a_0 \neq 0$, $a_k \neq 0$, *then it follows that* A_0, A_1 *are either both singular or both invertible.*)

(iv) *If the scheme converges, then for each* j *and for each sequence* $\{i_\ell\}_{\ell=1}^\infty$, *the product* (2.12) *tends to a multiple of the vector* $(1,\dots,1)$ *as* $n \to \infty$.

The singularity of the matrix A_0 is determined by the following result [GANT, p. 197].

THEOREM 2.5 (Orlando's formula). A_0 *is a singular matrix if and only if the polynomial* $A(z) = \sum_{i=0}^k a_i z^i$ *has a pair of opposite zeros* z *and* $-z$.

Positive schemes have the following property.

THEOREM 2.6. *Let* $\{a_i\}_{i=0}^k$, $k>1$, *be a positive mask satisfying* (2.4); *then*

(i) *The associated subdivision scheme converges*;

(ii) [MPIN]. *The refinable function* $E(t)$ *is positive on* $(0,k)$.

Now, we focus our attention on a Hurwitz sequence, $\{a_i\}_{i=0}^k$. The following theorem is a direct consequence of Theorems 2.3 and 2.6 (see also [GM], [Y]).

THEOREM 2.7. *Let* $\{a_i\}_{i=0}^k$ *be a Hurwitz sequence satisfying* (2.4), *and assume that* $A(z)$ *has an* m*-fold zero at* $z=-1$, $1 \le m < k$. *The scheme and its associated divided difference schemes up to order* $m-1$ *are convergent;* $\{a_i\}_{i=0}^k, \dots, \{a_i^{(m-1)}\}_{i=0}^{k-m+1}$ *are Hurwitz sequences.*

A key result here is that the bi-infinite matrix

$$A_{ij} = a_{i-2j} \tag{2.13}$$

is totally nonnegative if $\{a_i\}_{i=0}^k$ is a Hurwitz sequence (see [KEMP], [AS]).

Now applying (2.1) to the data at level n, $\{f_i^n\} \subset R$, it follows that

$$S^-\{f^{n+1}\} \subset S^-\{f^n\}, \tag{2.14}$$

and, consequently,

$$S^-\left\{\sum_i f_i^0 E(\cdot - i)\right\} \le S^-\{f_i^0\} \ . \tag{2.15}$$

This argument appears in [MP] for the particular case of a Polya frequency sequence (all roots of $\sum_{i=0}^k a_i z^i$ are real and negative).

3. Local linear independence. In the following let $\{a_i\}_{i=0}^k$ $(a_0 \neq 0\ ,\ a_k \neq 0)$ determine a convergent scheme, and let $E(t)$ be the associated refinable function that vanishes outside of the interval $(0,k)$. By Theorem 2.4 (iii) A_0 and A_1 are either both singular or both invertible.

DEFINITION 3.1. The functions $\{E(t-i)\}_{i=-\infty}^{\infty}$ are said to be locally linearly independent if $\{E(t-i)\}_{i=-(k-1)}^{0}$ are linearly independent on each subinterval of $[0,1]$.

LEMMA 3.2. *Assume that A_0 is invertible and the scheme converges*; *then the following statements are equivalent*:

(i) $\{E(t-i)\}_{i=-\infty}^{\infty}$ *are locally linearly independent*;

(ii) $\{E(t-i)\}_{i=-(k-1)}^{0}$ *are linearly independent on* [0,1].

Proof. We prove (ii) $\Rightarrow$ (i) (the converse is trivial). Assume in contradiction that $\sum_{i=-(k-1)}^{0} f_i^0 E(t-i) \not\equiv 0$ on [0,1], but there exist j and n such that

$$\sum_{i=-(k-1)}^{0} f_i^0 E(t-i) \equiv 0, \qquad t \in [\,j2^{-n},\ (j+1)2^{-n}\,].$$

By (2.11) it follows that

$$\sum_{i=j-(k-1)}^{j} f_i^n E(2^n t-i) \equiv 0, \qquad t \in [\,j2^{-n},\ (j+1)2^{-n}\,],$$

which implies that one of the following statements is true.

(a) $\{f_i^n\}_{i=j-(k-1)}^{j}$ is trivial.

(b) $\{E(2^n t-i)\}_{i=j-(k-1)}^{j}$ are linearly dependent on $[\,j2^{-n}, (j+1)2^{-n}\,]$.

Now, (b) is a contradiction to (ii) and (a), together with (2.12), is a contradiction to the nonsingularity of A_0 and A_1. □

THEOREM 3.3. *Let $\{a_i\}_{i=0}^k$ be a Hurwitz sequence satisfying* (2.4). *Then $\{E(t-i)\}$ are locally linearly independent.*

Proof. By Theorem 2.6 the scheme converges, and $\{E(t-i)\}_{i=-(k-1)}^{0}$ are positive functions on (0,1). Since all roots of $\sum_{i=0}^k a_i z^i$ are in the left half plane, then by Theorem 2.5 A_0 and A_1 are invertible. Thus, by Lemma 3.2 we discuss linear independence on $[0,1]$. Assume that $\sum_{i=-(k-1)}^{0} f_i^0 E(t-i) \equiv 0$ on $[0,1]$ and $\overline{f_0^0} = \{f_i^0\}_{i=-(k-1)}^{0}$ is nontrivial; then by the positivity of $\{E(t-i)\}$ it follows that $S^-(\overline{f_0^0}) > 0$. Also for each n and each $\{i_\ell\}_{\ell=0}^n$ we get by the nonsingularity of A_0, A_1 that $\prod_{\ell=1}^n A_{i_\ell}\overline{f_0^0}$ is nontrivial, and, consequently, $S^-(\prod_{\ell=1}^n A_{i_\ell}\overline{f_0^0}) > 0$. Thus, $S^-\{f_i^n\}$ tends to infinity, contradicting (2.14). □

The rest of this section is devoted to positive and interpolatory schemes, and we will not make use of these results in the following sections.

THEOREM 3.4. *Let $\{a_i\}_{i=0}^k$ be a positive mask satisfying* (2.4), *and let A_0 be invertible. Then the following statements are equivalent*:

(i) $\{E(t-i)\}$ *are locally linearly independent*;

(ii) *Each nontrivial common invariant subspace of A_0 and A_1 owns a nontrivial, nonnegative vector.*

(Note that at least one nontrivial common invariant subspace always exists, namely, $(1,1,\dots,1)$.)

Proof. As in the preceding proof the scheme converges, $\{E(t-i)\}_{i=-(k-1)}^{0}$ are positive on (0,1) and (i) is replaced by linear independence on $[0,1]$.

(ii) ⇒ (i) Let M be the space of all vectors $\overline{f_0^0}$ such that $\sum_i f_i^0 E(t-i) \equiv 0$ on $[0,1]$, and assume in contradiction that M is nontrivial. It is clear that M is a common invariant subspace of A_0, A_1, and by the positivity of $\{E(t-i)\}$ it follows that each nontrivial vector in M has at least one strong sign change, contradicting (ii).

(i) ⇒ (ii) Assume in contradiction that there exists a nontrivial common invariant subspace M such that each nontrivial vector in M has a strong sign change. Let $\overline{f_0^0} \in M$ be a nontrivial vector. Then all vectors of the form $\prod_{\ell=1}^{n} A_{i_\ell}\overline{f_0^0}$ are also in M, and also have a strong sign change. Since the scheme converges, $\lim_{\ell\to\infty}\prod_{\ell=1}^{\infty} A_{i_\ell}\overline{f_0^0}$ converges to a multiple of $(1,\ldots,1)$, which implies $\lim_{\ell\to\infty}\prod_{\ell=1}^{\infty} A_{i_\ell}\overline{f_0^0} = 0$, contradicting the linear independence. □

A scheme is said to be interpolatory if there exists $j \in Z$ such that $a_{2i+j} = \delta_{i,0}$. Obviously, $\{f_i^n\}$ is a subset of $\{f_i^{n+1}\}$, and if the scheme converges, then $E(i) = \delta_{i,j}$ for each $i \in Z$.

THEOREM 3.5. *Let $\{a_i\}_{i=0}^{k}$ determine a convergent interpolatory scheme, and assume that A_0 is invertible. Then $\{E(t-i)\}$ are locally linearly independent.*

Proof. Assume in contradiction that there exists a nontrivial $\{f_i^0\}_{i=-(k-1)}^{0}$ such that $\sum_{i=-(k-1)}^{0} f_i^0 E(t-i) \equiv 0$, on $[0,1]$. By the nonsingularity of A_0, A_1 it follows that for each n and each $\{i_\ell\}_{\ell=1}^{n}$, the vector $\overline{f_j^n} = \prod_{\ell=1}^{n} A_{i_\ell}\overline{f_0^0}$ is nontrivial. For a sufficiently large n there exists j such that the support of the functions $\{E(2^n t - i)\}_{i=j-(k-1)}^{j}$ is contained in $[0,1]$, and since the scheme is interpolatory, then $\{f_i^n\}_{i=j-(k-1)}^{j}$ are values of $\sum_{i=-(k-1)}^{0} f_i^0 E(t-i)$. Hence, there exists t such that $\sum_{i=-(k-1)}^{0} f_i^0 E(t-i) \neq 0$, and the contradiction is obtained. □

Remark 3.6. The four-point scheme (see, e.g., [D]) depends on a positive tension parameter w and is determined by

(3.1)
$$a_0 = -w, \quad a_1 = 0, \quad a_2 = \tfrac{1}{2} + w, \quad a_3 = 1, \quad a_4 = \tfrac{1}{2} + w, \quad a_5 = 0, \quad a_6 = -w.$$

The scheme is known to converge for $|w| < 0.5$. The polynomial $\sum_{i=0}^{k} a_i z^i$ has no pair of opposite zeros; thus, by Theorem 3.5 the $\{E(t-i)\}$ are locally linearly independent.

4. An alternative analysis of Lagrange interpolation.

Let $\{a_i\}_{i=0}^{k}$ be a Hurwitz sequence satisfying (2.4). Then $E(t)$ is continuous, and the following properties of $E(t)$ have already been discussed:

$$E(t) = 0, \qquad t \le 0\ , \quad t \ge k, \tag{4.1}$$

$$E(t) > 0, \qquad 0 < t < k, \tag{4.2}$$

$$S^-\left(\sum_i \alpha_i E(\cdot - i)\right) \le S^-\{\alpha_i\}, \tag{4.3}$$

$$\{E(\cdot - i)\},\ i \in Z \text{ are locally linearly independent.} \tag{4.4}$$

Only these properties of $E(t)$ play a role in this section.

In the following, let $\{\ell_\nu\}_{\nu=0}^{r}$ and $\{x_{\ell_\nu}\}_{\nu=0}^{r}$, $r \ge 0$, be strictly increasing sequences, and we define the set of points I by

$$I = \bigcup_{\nu=0}^{r} \text{supp}\,(E(.-\ell_\nu)) = \bigcup_{\nu=0}^{r} (\ell_\nu,\ \ell_\nu + k). \tag{4.5}$$

THEOREM 4.1.

$$D\begin{pmatrix} x_{\ell_0} & \cdots & x_{\ell_r} \\ E(\cdot - \ell_0) & \cdots & E(\cdot - \ell_r) \end{pmatrix} \geq 0 \quad \forall_\nu, \quad x_{\ell_\nu} < x_{\ell_{\nu+1}}. \tag{4.6}$$

Proof. We assume that $x_{\ell_\nu} \in I$ for all $_\nu$, since otherwise the determinant vanishes. For the particular sequence $x_{\ell_\nu} = \ell_\nu + \frac{1}{2}$ it is clear that the matrix (4.6) is lower triangular, and by (4.2) the diagonal is positive.

Now, we are under the following conditions:

(i) $\{E(t - \ell_\nu)\}_{\nu=0}^r$ is a sequence of continuous linearly independent functions on I;

(ii) There exists a sequence $\{x_{\ell_\nu}\}_{\nu=0}^r$ such that (4.6) is positive;

(iii) On I, $S^-(\sum_{\nu=0}^r \alpha_{\ell_\nu} E(\cdot - \ell_\nu)) \leq S^-\{\alpha_{\ell_\nu}\}_{\nu=0}^r$ for each sequence $\{\alpha_{\ell_\nu}\}_{\nu=0}^r$;

(i) and (iii), together with [SCH] Theorem 2.39, imply that either $\{E(\cdot - \ell_\nu)\}_{\nu=0}^r$ or $\{E(\cdot - \ell_0), \ldots, E(\cdot - \ell_{r-1}), -E(\cdot - \ell_r)\}$ is a WT-system on I. Hence, (4.6) is either nonnegative or nonpositive, and the proof is completed by (ii). □

THEOREM 4.2.

$$D\begin{pmatrix} x_{\ell_0} & \cdots & x_{\ell_r} \\ E(\cdot - \ell_0) & \cdots & E(\cdot - \ell_r) \end{pmatrix} > 0, \qquad x_{\ell_\nu} < x_{\ell_{\nu+1}} \tag{4.7}$$

if and only if

$$\forall \nu, \quad x_{\ell_\nu} \in \operatorname{supp}(E(\cdot - \ell_\nu)) = (\ell_\nu \ , \ \ell_\nu + k). \tag{4.8}$$

Proof. First, we prove that (4.7) ⇒ (4.8). Assume in contradiction that (4.8) is violated while (4.7) still holds. If the violation occurs at $\nu = r$, i.e., $x_{\ell_r} \notin \operatorname{supp}(E(\cdot - \ell_r))$, then either the last row of (4.7) or the last column of (4.7) vanish, contradicting the positivity of the determinant. Hence, the violation occurs at $\nu = q$, where $q < r$. By Theorem 4.1 it follows that (4.7) is a totally nonnegative matrix, and by [GANT, p. 100] we get

(4.9)

$$\begin{aligned} & D\begin{pmatrix} x_{\ell_0} & \cdots & x_{\ell_r} \\ E(\cdot - \ell_0) & \cdots & E(\cdot - \ell_r) \end{pmatrix} \\ & \quad \leq D\begin{pmatrix} x_{\ell_0} & \cdots & x_{\ell_q} \\ E(\cdot - \ell_0) & \cdots & E(\cdot - \ell_q) \end{pmatrix} \cdot D\begin{pmatrix} x_{\ell_{q+1}} & \cdots & x_{\ell_r} \\ E(\cdot - x_{\ell_{q+1}}) & \cdots & E(\cdot - x_{\ell_r}) \end{pmatrix}. \end{aligned}$$

Now by a preceding argument we get

$$D\begin{pmatrix} x_{\ell_0} & \cdots & x_{\ell_q} \\ E(\cdot - \ell_0) & \cdots & E(\cdot - \ell_q) \end{pmatrix} = 0, \tag{4.10}$$

and together with (4.9) it follows that (4.7) vanishes.

In the following we prove (4.8) ⇒ (4.7). The case $r = 0$ is obviously true, and we assume by induction that it is true for each sequence $\{x_{\ell_\nu}\}_{\nu=0}^j$, $j < r$. Assume in contradiction that (4.8) holds, and there exists $f(t)$ such that

$$f(t) = \sum_{\nu=0}^r \alpha_{\ell_\nu} E(t - \ell_\nu), \quad f(x_{\ell_\nu}) = 0 \quad \forall \nu, \quad \{\alpha_{\ell_\nu}\} \text{ nontrivial.} \tag{4.11}$$

First note that

$$\alpha_{\ell_\nu} \neq 0, \qquad 0 \leq \nu \leq r \tag{4.12}$$

since if $\alpha_{\ell_\nu} = 0$, then by deleting row ℓ_ν and column ℓ_ν the induction hypothesis is contradicted. In view of the local linear independence, $f(t)$ does not vanish on any subinterval contained in I.

Now we are under the following conditions:

(i) $\{E(t-\ell_\nu)\}_{\nu=0}^r$ is a WT-system of order $r+1$ on I;

(ii) $f(t)$ has $r+1$ zeros on I, namely, $\{x_{\ell_\nu}\}_{\nu=0}^r$;

(iii) The zeros of f are essential (since $E_{\ell_\nu}(x_{\ell_\nu}) \neq 0$; see [SCH, Def. 2.43]);

By [SCH, Thm. 2.45] it follows that $f(t)$ is vanishing on a subinterval of I, and the proof is complete. □

5. Hermite interpolation problem. In the following we assume that $\{a_i\}_{i=0}^k$ is a Hurwitz sequence satisfying (2.4), and that the characteristic polynomial $\sum_{i=0}^k a_i z^i$ has an $(m+1)$-fold zero at $z=-1$, $1 < m+1 < k$. By the results of §2 it follows that $E \in C^m$.

In general the Hermite interpolation problem of order m involves derivatives up to order m, and the interpolation points $\{x_{\ell_\nu}\}_{\nu=0}^r$ satisfy

$$x_{\ell_\nu} \le x_{\ell_{\nu+1}}, \qquad x_{\ell_\nu} < x_{\ell_{\nu+m+1}}, \tag{5.1}$$

where $\{\ell_\nu\}$ is a strictly increasing sequence of integers. The matrix we discuss is

$$M\begin{pmatrix} x_{\ell_0} & \cdots & x_{\ell_r} \\ E(\cdot-\ell_0) & \cdots & E(\cdot-\ell_r)\end{pmatrix}_{i,j} = \frac{\partial}{\partial^{d_i}} E(x_{\ell_i}-\ell_j), \tag{5.2}$$

$$i,j = 0,\dots,r, \quad d_i = \max\{q : x_{\ell_i} = \cdots x_{\ell_{i-q}}\}.$$

In words, if $x_{\ell_\nu} = \cdots = x_{\ell_{\nu-q}} \neq x_{\ell_{\nu-q-1}}$, then the $\nu-q$ row contains the values of $\{E(x_{\ell_\nu}-\ell_\gamma)\}_{\gamma=0}^r$, the $\nu-q+1$ row contains the values of $\{E'(x_{\ell_\nu}-\ell_\gamma)\}_{\gamma=0}^r$, and the ν row contains the values of $\{E^{(q)}(x_{\ell_\nu}-\ell_\gamma)\}_{\gamma=0}^r$.

The following theorem is an immediate result of Theorem 4.1 and [SCH, Lemma 2.9].

THEOREM 5.1. *For each* $\{x_{\ell_\nu}\}_{\nu=0}^r$, $x_{\ell_\nu} \le x_{\ell_{\nu+1}}$, $x_{\ell_\nu} < x_{\ell_{\nu+m+1}}$, *we then have*

$$D\begin{pmatrix} x_{\ell_0} & \cdots & x_{\ell_r} \\ E(\cdot-\ell_0) & \cdots & E(\cdot-\ell_r)\end{pmatrix} \ge 0. \tag{5.3}$$

The following theorem provides a necessary condition for nonsingularity of (5.3).

THEOREM 5.2. *For each* $\{x_{\ell_\nu}\}_{\nu=0}^r$, $x_{\ell_\nu} \le x_{\ell_{\nu+1}}$, $x_{\ell_\nu} < x_{\ell_{\nu+m+1}}$, *we then have*

$$D\begin{pmatrix} x_{\ell_0} & \cdots & x_{\ell_r} \\ E(\cdot-\ell_0) & \cdots & E(\cdot-\ell_r)\end{pmatrix} > 0 \tag{5.4}$$

only if

$$x_{\ell_\nu} \in \operatorname{supp}\,(E(\cdot-\ell_\nu)) \quad \forall\nu, \quad 0 \le \nu \le r. \tag{5.5}$$

Proof. Assume in contradiction that (5.5) is violated, i.e., there exists ν such that $x_{\ell_\nu} \notin \operatorname{supp}(E(\cdot-\ell_\nu))$. Substituting all the derivative values appearing in the matrix (5.4) by discrete divided differences according to the formula

$$\frac{\partial}{\partial^q} E(x_{\ell_i}-\ell_j) := q!\ [x_{\ell_i}, x_{\ell_i+h}, \dots, x_{\ell_i+qh}]E(\cdot-\ell_j), \tag{5.6}$$

the modified determinant is clearly zero for a sufficiently small h. Hence, (5.4) vanishes and the proof is completed. □

Now, we restrict ourselves to consecutive translates, and under this assumption Theorem 5.3 is a converse for Theorem 5.2. Originally, Theorem 5.3 was proved for the case $m = 1$, and its generalization to the current form is due to Mark Kon.

THEOREM 5.3. *Let* $\{x_\ell\}_{\ell=0}^r$ *satisfy*

$$x_\ell \le x_{\ell+1}, \quad x_\ell < x_{\ell+m+1}, \quad x_\ell \in \operatorname{supp}\,(E(\cdot - \ell)). \tag{5.7}$$

Then

$$D\begin{pmatrix} x_0 & \cdots & x_r \\ E(\cdot) & \cdots & E(\cdot - r) \end{pmatrix} > 0. \tag{5.8}$$

Proof. The case $r = 0$ is obvious, and we assume by induction that (5.7) $\Rightarrow$ (5.8) for each sequence $\{x_\ell\}_{\ell=0}^j$, $j < r$. Assume in contradiction that there exists $\{x_\ell\}_{\ell=0}^r$ satisfying (5.7), where (5.8) is vanishing. Hence, there exists $f(t)$ such that

$$f(t) = \sum_{\ell=0}^{r} \alpha_\ell E(t - \ell), \qquad \{\alpha_\ell\}_{\ell=0}^r \text{ nontrivial} \tag{5.9}$$

with

$$f(x_\ell) = 0 \quad \forall \ell, \quad 0 \le \ell \le r \tag{5.10}$$

and

$$\frac{\partial^j}{\partial t^j} f(x_\ell) = 0 \quad \text{if} \quad x_\ell = x_{\ell-j}, \quad 1 < j \le m\ . \tag{5.11}$$

First observe that $\alpha_0 \ne 0$, since otherwise

$$D\begin{pmatrix} x_1 & \cdots & x_r \\ E(\cdot - 1) & \cdots & E(\cdot - r) \end{pmatrix} = 0 \tag{5.12}$$

in both cases, $x_0 < x_1$ and $x_0 = x_1$, contradicting the induction hypothesis. By the same argument, $\alpha_r \ne 0$. Next observe that $f(t)$ does not vanish on any subinterval of

$$I = \bigcup_{\ell=0}^{r} \operatorname{supp}\,(E(\cdot - \ell)) = (0, r + k). \tag{5.13}$$

It is true since by the local linear independence and the previous observation it follows that if f vanishes on a subinterval, then there exists a trivial sequence $\{\alpha_\ell\}_{\ell=q}^{q+k-1}$ with $0 < q$, $q + k - 1 < r$. Since the support of E is of size k, then

$$D\begin{pmatrix} x_0 & \cdots & x_{q-1} \\ E(\cdot) & \cdots & E(\cdot - (q - 1)) \end{pmatrix} = 0, \tag{5.14}$$

contradicting the induction hypothesis.

Now, we count the zeros of f' on I. Let j be the number of points x_ℓ with $x_\ell \ne x_{\ell-1}$. Then f' has $r + 2$ zeros on I according to the following count:

(i) $r + 1 - j$ zeros follow from the multiplicities in $\{x_\ell\}_{\ell=0}^r$;

(ii) $j-1$ zeros follow from Rolle's theorem;

(iii) Two zeros near the boundaries of I follow also from Rolle's theorem since $f(0)=f(k+r)=0$.

By the results of §2 it is known that

$$E'(t)=F(t)-F(t-1), \tag{5.15}$$

where $F(t)$ is the refinable function of the divided difference scheme, which is also determined by a Hurwitz sequence, and $F(t)$ is supported on $(0,k-1)$. Hence $f'(t)$ is a function of the form

$$f'(t)=\sum_{\ell=0}^{r+1}\beta_\ell F(t-\ell), \tag{5.16}$$

and due to the above construction its zeros $\{y_i\}_{i=0}^{r+1}$ satisfy

$$0<y_0\le\cdots\le y_{r+1}<(r+1)+(k-1)\ , \qquad y_\ell<y_{\ell+m}. \tag{5.17}$$

The proof is complete by the following inductive argument on m. This theorem is true for $m=0$ (by the preceding section), and assume it is true for $0,\dots,m-1$. By Lemma 5.4 it follows that there exists some i such that $f'(t)$ vanishes on a subinterval containing y_i and y_{i+1}. Now, since one of the zeros (or both) is a zero of f, then f also vanishes on this subinterval, a contradiction. □

The following lemma is needed for the preceding proof.

LEMMA 5.4. *Assume that Theorem* 5.3 *holds for* $m=0,\dots,n$, *and let* $f(t)=\sum_{\ell=0}^{r}\alpha_\ell E(t-\ell)$, $1\le r$ *vanish at* $y_0,\dots,y_r$. *If*

$$0<y_0\le\cdots\le y_r<r+k, \qquad y_\ell<y_{\ell+n+1} \quad \forall\ell\ , \tag{5.18}$$

then there exists i *such that* f *vanishes on some interval containing* y_i,y_{i+1}.

Proof. By induction on r, the proof of the case $r=1$ is easy, and we assume the hypothesis is true for $1,\dots,r-1$.

First observe that there exists some q such that

$$y_q\notin(q,q+k)\ , \tag{5.19}$$

since otherwise $f\equiv0$, hence we assume $y_q\le q$. (The case $y_q\ge q+k$ is similar.) Define

$$g(t)=\sum_{\ell=0}^{q-1}\alpha_\ell E(t-\ell); \tag{5.20}$$

then $g=f$ on $(0,q]$. By induction hypothesis g vanishes on a subinterval of $(0,q]$ containing y_i,y_{i+1}, $i<q$, and the proof is completed. □

Acknowledgments. I would like to thank N. Dyn for useful discussions and the referees for their helpful comments. In particular, I would like to thank Mark Kon for the generalization of Theorem 5.3.

REFERENCES

[AS] B. A. ASNER (1970), *On the total non-negativity of Hurwitz matrix*, SIAM J. Appl. Math., 18, pp. 407–414.

[CDM] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli (1991), *Stationary subdivision*, Mem. Amer. Math. Soc., 453, pp. 1–186.

[CW] C. K. Chui and J. Z. Wang (1990), *A general framework of compactly supported splines and wavelets*, CAT Report 219, Texas A&M University, College Station, TX.

[DAUB] I. Daubechies (1988), *Orthonormal bases of compactly supported wavelets*, Pure Appl. Math., 41, pp. 909–996.

[DL] I. Daubechies and C. Lagarias (1991), *Two scale difference equations* I *and* II, SIAM J. Math. Anal., 22, pp. 1388–1410; 23 (1992), pp. 1031–1079.

[D] N. Dyn (1991), *Subdivision schemes in computer aided graphic design*, in Advances in Numerical Analysis II, Subdivision Algorithms and Radial Functions, W. A. Light, ed., Oxford University Press, London, pp. 36–104.

[DGL] N. Dyn, J.A. Gregory, and D. Levin (1991), *Analysis of linear binary subdivision schemes for curve design*, Constr. Approx., 7, pp. 127–147.

[GANT] F. R. Gantmacher (1959), *The theory of matrices, Vol.* II, Chelsea, New York.

[G] T. N. T. Goodman (1989), *Shape preserving representations*, in Mathematical Methods in CAGD, T. Lyche and L. L. Schumaker, eds., pp. 333–351.

[GM] T. N. T. Goodman and C. A. Micchelli (1992), *On refinement equations determined by Polya frequency sequences*, SIAM J. Math. Anal., 23, pp. 766–784.

[KEMP] J. H. B. Kemperman (1982), *A Hurwitz matrix is totally positive*, SIAM J. Math. Anal., 13, pp. 331–341.

[M] C. A. Micchelli, *Using the refinement equation for the construction of pre-wavelets*, Numer. Algorithms, to appear.

[MPIN] C. A. Micchelli and A. Pinkus (1991), *Descartes systems from corner cutting*, Constr. Approx., 7, pp. 161–194.

[MP] C. A. Micchelli and H. Prautzch (1987), *Refinement and subdivision for spaces of integer translates of a compactly supported function*, in Numerical Analysis, D. F. Griffiths and G. A. Watson, eds., pp. 192–222.

[SCH] L. L. Schumaker (1981), *Spline functions basic theory*, John Wiley, New York.

[Y] I. Yad-Shalom, *Monotonicity preserving subdivision schemes*, J. Approx. Theory, to appear.

SIAM J. MATH. ANAL.
Vol. 24, No. 1, pp. 246–262, January 1993

A CLASS OF BASES IN L^2 FOR THE SPARSE REPRESENTATION OF INTEGRAL OPERATORS*

BRADLEY K. ALPERT[†]

Abstract. A class of *multiwavelet* bases for L^2 is constructed with the property that a variety of integral operators is represented in these bases as sparse matrices, to high precision. In particular, an integral operator $\mathcal{K}$ whose kernel is smooth except along a finite number of singular bands has a sparse representation. In addition, the inverse operator $(I - \mathcal{K})^{-1}$ appearing in the solution of a second-kind integral equation involving $\mathcal{K}$ is often sparse in the new bases. The result is an order $O(n \log^2 n)$ algorithm for numerical solution of a large class of second-kind integral equations.

Key words. wavelets, integral equations, sparse matrices

AMS(MOS) subject classifications. 42C15, 45L10, 65R10, 65R20

Introduction. Families of functions $h_{a,b}$,

$$h_{a,b}(x) = |a|^{-1/2}\, h\left(\frac{x-b}{a}\right), \qquad a, b \in \mathbf{R}, \quad a \neq 0,$$

derived from a single function h by dilation and translation, which form a basis for $L^2(\mathbf{R})$, are known as *wavelets* (Grossman and Morlet [9]). In recent years, these families have received study by many authors, resulting in constructions with a variety of properties. Meyer [11] constructed orthonormal wavelets for which $h \in C^\infty(\mathbf{R})$. Daubechies [6] constructed compactly supported wavelets with $h \in C^k(\mathbf{R})$ for arbitrary k, and [6] gives an overview and synthesis of the field.

Beylkin, Coifman, and Rokhlin [4] develop the connection between wavelets and recent fast numerical algorithms devised by Rokhlin and other authors [3], [8], [14], [15]. These algorithms exploit analytical properties of specific linear operators to achieve, in each case, fast application of an operator to an arbitrary function. The operator and function are discretized to a matrix and vector; in the discrete representation a full $n \times n$-matrix is applied to a vector of length n in order $O(n)$ operations, as opposed to order $O(n^2)$ operations for naive matrix-vector multiplication. Each algorithm depends on "local smoothness" of the underlying operator. In particular, each algorithm may be viewed as the division of the operator matrix into order $O(n)$ square submatrices, each approximated by a matrix of low rank, followed by the fast application of the submatrices to the function vector.

In [4] it is observed that these numerical algorithms can be generalized by a technique in which the underlying operator is represented in a basis of wavelets. Discretization (i.e., truncation of the operator expansion) then results in an operator matrix that is approximated by a sparse matrix. The characteristics of the wavelets bases which lead to a sparse matrix representation are that

(1) The basis functions are orthogonal to low-order polynomials (have vanishing moments); and

* Received by the editors December 26, 1990; accepted for publication (in revised form) May 5, 1992.

[†] Lawrence Berkeley Laboratory and University of California, Berkeley, California. This research was supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy under contract DE-AC03-76SF00098, Office of Naval Research grant N00014-86-0310, Defense Advanced Research Projects Agency grant DMS-9012751, and IBM grant P00038437. Present address, National Institute of Standards and Technology, Boulder, Colorado 80303 (alpert@bldr.nist.gov).

(2) Most basis functions have small intervals of support.

An integral operator whose kernel is a smooth, nonoscillatory, function of its arguments over most of their range (and, therefore, can be approximated locally by low-order polynomials) will have negligible projection on most basis functions.

One difficulty of using wavelets bases for the representation of integral operators is that they do not form a basis for functions on a finite interval. Wavelet basis functions overlap in such a way that either the interval must be extended, a periodization must be performed, or the basis functions at the interval ends must be modified. In [4] the integrand is treated as periodic, with some loss of sparsity. In [13] Meyer showed how the basis functions overlapping the interval ends can be truncated and reorthogonalized to obtain a basis on the finite interval.

A second difficulty of using wavelets for the representation of integral operators is that projection onto the basis functions requires appropriate integration quadratures (as is true with other bases). The order of convergence of the quadratures determines the order of the numerical method as a whole. The difficulty is that quadratures must be employed for each element of the resulting matrix, leading to potentially high cost. On the other hand, use of the Nyström method, in which the interval is discretized into n points and the integral at each point is approximated by a quadrature, requires the application of quadratures only n times.

In this paper we construct a class of wavelet-like bases, which we call *multiwavelet* bases, which lead to the sparse representation of smooth integral operators on a finite interval. For each basis, the interval is recursively bisected; the basis functions on a given scale are supported on the dyadic subintervals of a particular size. Out of this class of bases, different bases differ in the number of basis functions supported on each subinterval, and this number corresponds to the order of convergence of expansions of C^∞ functions. The lack of overlap of the basis functions on a single scale eliminates the first difficulty (mentioned above) of using wavelets for the representation of integral operators. The second difficulty is eliminated by the construction of a discrete counterpart to the bases developed here. The latter construction is described in [2]. A principal advantage of the present construction is its simplicity.

In §1, we construct multiwavelet bases, and in §2 we prove that the representations in these bases of certain integral operators are sparse, to high precision. In §3 we give several numerical examples of the bases and the solution of second-kind integral equations and conclude with a discussion.

1. Multiwavelet bases.

1.1. The one-dimensional construction.

We construct a class of bases for $L^2(\mathbf{R})$ that can be readily revised to bases for $L^2[0,1]$. Each basis is comprised of dilates and translates of a finite set of functions $h_1, \dots, h_k$. In particular, these bases consist of orthonormal systems

$$(1) \qquad h^n_{j,m}(x) = 2^{m/2}\, h_j(2^m x - n), \qquad j = 1, \dots, k;\ m, n \in \mathbf{Z},$$

where the functions $h_1, \dots, h_k$ are piecewise polynomial, vanish outside the interval $[0,1]$, and are orthogonal to low-order polynomials (have vanishing moments),

$$(2) \qquad \int_0^1 h_j(x)\, x^i\, dx = 0, \qquad i = 0, 1, \dots, k-1.$$

We first restrict our attention to the finite interval $[0,1] \subset \mathbf{R}$, and we construct a basis for $L^2[0,1]$. We employ the multiresolution analysis framework developed by

Mallat [10] and Meyer [12], and discussed at length by Daubechies [6]. We suppose that k is a positive integer, and for $m = 0, 1, 2, \dots$ we define a space S^k_m of piecewise polynomial functions,

$$(3) \quad S^k_m = \{f : \begin{array}{l}\text{the restriction of } f \text{ to the interval } (2^{-m}n, 2^{-m}(n+1)) \text{ is}\\ \text{a polynomial of degree less than } k, \text{ for } n = 0, \dots, 2^m - 1,\\ \text{and } f \text{ vanishes elsewhere}\}.\end{array}$$

It is apparent that the space S^k_m has dimension $2^m k$ and

$$S^k_0 \subset S^k_1 \subset \cdots \subset S^k_m \subset \cdots.$$

For $m = 0, 1, 2, \dots$ we define the $2^m k$-dimensional space R^k_m to be the orthogonal complement of S^k_m in S^k_{m+1},

$$S^k_m \oplus R^k_m = S^k_{m+1}, \qquad R^k_m \perp S^k_m;$$

so we inductively obtain the decomposition

$$(4) \qquad S^k_m = S^k_0 \oplus R^k_0 \oplus R^k_1 \oplus \cdots \oplus R^k_{m-1}.$$

Suppose that functions $h_1, \dots, h_k : \mathbf{R} \to \mathbf{R}$ form an orthogonal basis for R^k_0. Since R^k_0 is orthogonal to S^k_0, the first k moments of $h_1, \dots, h_k$ vanish,

$$\int_0^1 h_j(x)\, x^i\, dx = 0, \qquad i = 0, 1, \dots, k-1.$$

The $2k$-dimensional space R^k_1 is spanned by the $2k$ orthogonal functions $h_1(2x), \dots, h_k(2x)$, $h_1(2x-1), \dots, h_k(2x-1)$, of which k are supported on the interval $[0, \frac{1}{2}]$ and k on $[\frac{1}{2}, 1]$. In general, the space R^k_m is spanned by $2^m k$ functions obtained from $h_1, \dots, h_k$ by translation and dilation. There is some freedom in choosing the functions $h_1, \dots, h_k$ within the constraint that they be orthogonal; by requiring normality and additional vanishing moments, we specify them uniquely, up to sign. The remainder of this subsection is devoted to the explicit construction of $h_1, \dots, h_k$; in the following sections we exploit only the property that $h_1, \dots, h_k$ form an orthonormal basis for R^k_0.

In preparation for the definition of $h_1, \dots, h_k$, we construct the k functions $f_1, \dots, f_k : \mathbf{R} \to \mathbf{R}$, supported on the interval $[-1, 1]$, with the following properties:

(1) The restriction of f_i to the interval $(0, 1)$ is a polynomial of degree $k - 1$;

(2) The function f_i is extended to the interval $(-1, 0)$ as an even or odd function according to the parity of $i + k - 1$;

(3) The functions $f_1, \dots, f_k$ satisfy the following orthogonality and normality conditions:

$$\int_{-1}^1 f_i(x)\, f_j(x)\, dx \equiv \langle f_i, f_j \rangle = \delta_{ij}, \qquad i, j = 1, \dots, k;$$

(4) The function f_j has vanishing moments,

$$\int_{-1}^1 f_j(x)\, x^i\, dx = 0, \qquad i = 0, 1, \dots, j + k - 2.$$

Properties 1 and 2 imply that there are k^2 polynomial coefficients that determine the functions $f_1, \ldots, f_k$, while properties 3 and 4 provide k^2 (nontrivial) constraints. It turns out that the equations uncouple to give k nonsingular linear systems that may be solved to obtain the coefficients, yielding the functions uniquely (up to sign). Rather than prove that these systems are nonsingular, however, we now determine $f_1, \ldots, f_k$ constructively.

We start with $2k$ functions that span the space of functions that are polynomials of degree less than k on the interval $(0,1)$ and on $(-1,0)$, then orthogonalize k of them, first to the functions $1, x, \ldots, x^{k-1}$, then to the functions $x^k, x^{k+1}, \ldots, x^{2k-1}$, and finally among themselves. We define $f_1^1, f_2^1, \ldots, f_k^1$ by the formula

$$f_j^1(x) = \begin{cases} x^{j-1}, & x \in (0,1), \\ -x^{j-1}, & x \in (-1,0), \\ 0, & \text{otherwise}, \end{cases}$$

and note that the $2k$ functions $1, x, \ldots, x^{k-1}, f_1^1, f_2^1, \ldots, f_k^1$ are linearly independent, hence span the space of functions that are polynomials of degree less than k on $(0,1)$ and on $(-1,0)$.

(1) By the Gram–Schmidt process we orthogonalize f_j^1 with respect to $1, x, \ldots, x^{k-1}$, to obtain f_j^2, for $j = 1, \ldots, k$. This orthogonality is preserved by the remaining orthogonalizations, which only produce linear combinations of the f_j^2.

(2) The next sequence of steps yields $k-1$ functions orthogonal to x^k, of which $k-2$ functions are orthogonal to x^{k+1}, and so forth, down to one function which is orthogonal to x^{2k-2}. First, if at least one of f_j^2 is not orthogonal to x^k, we reorder the functions so that it appears first, $\langle f_1^2, x^k \rangle \neq 0$. We then define $f_j^3 = f_j^2 - a_j \cdot f_0^2$ where a_j is chosen so $\langle f_j^3, x^k \rangle = 0$ for $j = 2, \ldots, k$, achieving the desired orthogonality to x^k. Similarly, we orthogonalize to $x^{k+1}, \ldots, x^{2k-2}$, each in turn, to obtain $f_1^2, f_2^3, f_3^4, \ldots, f_k^{k+1}$ such that $\langle f_j^{j+1}, x^i \rangle = 0$ for $i \leq j + k - 2$.

(3) Finally, we do Gram–Schmidt orthogonalization on $f_k^{k+1}, f_{k-1}^k, \ldots, f_1^2$, in that order, and normalize to obtain $f_k, f_{k-1}, \ldots, f_1$.

It is readily seen that the f_j satisfy properties (1)–(4) of the previous paragraph. Defining $h_1, \ldots, h_k : \mathbf{R} \to \mathbf{R}$ by the formula

$$h_i(x) = 2^{1/2} \, f_i(2x-1), \qquad i = 1, \ldots, k,$$

we obtain the equality

$$R_0^k = \text{linear span } \{h_i : \ i = 1, \ldots, k\},$$

and, more generally,

$$R_m^k = \text{linear span } \{h_{j,m}^n : \ h_{j,m}^n(x) = 2^{m/2} \, h_j(2^m x - n), \quad j = 1, \ldots, k; \ n = 0, \ldots, 2^m - 1\}. \tag{5}$$

We will show next that dilates and translates of the piecewise polynomial functions $h_1, \ldots, h_k$ form an orthonormal basis for $L^2(\mathbf{R})$. Furthermore, a subset of these dilates and translates, combined with a basis for S_0^k, forms a basis for $L^2[0,1]$.

1.2. Completeness of one-dimensional construction. We define the space S^k to be the union of the S_m^k, given by the formula

$$S^k = \bigcup_{m=0}^{\infty} S_m^k, \tag{6}$$

and observe that $\overline{S^k} = L^2[0,1]$. In particular, S^k contains the Haar basis for $L^2[0,1]$, consisting of functions piecewise constant on each of the subintervals $(2^{-m}n, 2^{-m}(n+1))$. Here the closure $\overline{S^k}$ is defined with respect to the L^2-norm,

$$\|f\| = \langle f, f\rangle^{1/2},$$

where the inner product $\langle f, g\rangle$ is defined by the formula

$$\langle f, g\rangle = \int_0^1 f(x)\, g(x)\, dx.$$

We let $\{u_1, \ldots, u_k\}$ denote an orthonormal basis for S_0^k; in view of (4), (5), and (6), the orthonormal system

$$\begin{aligned} B_k = {} & \{u_j : j = 1, \ldots, k\} \\ & \cup \{h_{j,m}^n : j = 1, \ldots, k;\ m = 0, 1, 2, \ldots;\ n = 0, \ldots, 2^m - 1\} \end{aligned}$$

spans $L^2[0,1]$; we refer to B_k as the *multiwavelet basis of order* k for $L^2[0,1]$.

Now we construct a basis for $L^2(\mathbf{R})$ by defining, for $m \in \mathbf{Z}$, the space $\tilde{S}_m^k$ by the formula

$$\begin{aligned} \tilde{S}_m^k = \{f : {} & \text{the restriction of } f \text{ to the interval } (2^{-m}n, 2^{-m}(n+1)) \text{ is} \\ & \text{a polynomial of degree less than } k, \text{ for } n \in \mathbf{Z}\} \end{aligned}$$

and observing that the space $\tilde{S}_{m+1}^k \backslash \tilde{S}_m^k$ is spanned by the orthonormal set

$$\{h_{j,m}^n :\ h_{j,m}^n(x) = 2^{m/2}\, h_j(2^m x - n),\ j = 1, \ldots, k;\ n \in \mathbf{Z}\}.$$

Thus $L^2(\mathbf{R})$, which is contained in $\overline{\bigcup_m \tilde{S}_m^k}$, has an orthonormal basis

$$\{h_{j,m}^n :\ j = 1, \ldots, k;\ m, n \in \mathbf{Z}\}.$$

1.3. Construction in multiple dimensions. The construction of bases for $L^2[0,1]$ and $L^2(\mathbf{R})$ can be extended to certain other function spaces, including $L^2[a,b]^d$ and $L^2(\mathbf{R}^d)$, for any positive integer d. We now outline this extension by giving the basis for $L^2[0,1]^2$, which is illustrative of the construction for any finite-dimensional space. We define the space $S_m^{k,2}$ by the formula

$$S_m^{k,2} = S_m^k \times S_m^k, \qquad m = 0, 1, 2, \ldots,$$

where S_m^k is defined by (3). We further define $R_m^{k,2}$ to be the orthogonal complement of $S_m^{k,2}$ in $S_{m+1}^{k,2}$,

$$S_m^{k,2} \oplus R_m^{k,2} = S_{m+1}^{k,2}, \qquad R_m^{k,2} \perp S_m^{k,2}.$$

Then $R_0^{k,2}$ is the space spanned by the orthonormal basis

$$\{u_i(x) h_j(y),\ h_i(x) u_j(y),\ h_i(x) h_j(y) :\ i, j = 1, \ldots, k\}.$$

Among these $3k^2$ basis elements each element $v(x,y)$ has no projection on low-order polynomials,

$$\int_0^1 \int_0^1 v(x,y)\, x^i\, y^j\, dx\, dy = 0, \qquad i, j = 0, 1, \ldots, k-1.$$

The space $R_m^{k,2}$ is spanned by dilations and translations of the $v(x,y)$ and the basis of $L^2[0,1]^2$ consists of these functions and the low-order polynomials $\{u_i(x) u_j(y) : i, j = 1, \ldots, k\}$.

1.4. Convergence of the multiwavelet bases. For a function $f \in L^2[0,1]$, a positive integer k, and $m = 0,1,2,\ldots$, we define the orthogonal projection $Q^k_m f$ of f onto S^k_m by the formula

$$(Q^k_m f)(x) = \sum_{j,n} \langle f, u^n_{j,m} \rangle \cdot u^n_{j,m}(x),$$

where $\{u^n_{j,m}\}$ is an orthonormal basis for S^k_m. The projections $Q^k_m f$ converge (in the mean) to f as $m \to \infty$. If the function f is several times differentiable, we can bound the error, as established by the following lemma.

LEMMA 1.1. *Suppose that the function $f : [0,1] \to \mathbf{R}$ is k times continuously differentiable, $f \in C^k[0,1]$. Then $Q^k_m f$ approximates f with mean error bounded as follows*:

$$\|Q^k_m f - f\| \le 2^{-mk} \frac{2}{4^k k!} \sup_{x\in[0,1]} |f^{(k)}(x)|. \tag{7}$$

Proof. We divide the interval $[0,1]$ into subintervals on which $Q^k_m f$ is a polynomial; the restriction of $Q^k_m f$ to one such subinterval $I_{m,n}$ is the polynomial of degree less than k that approximates f with minimum mean error. We then use the maximum error estimate for the polynomial which interpolates f at Chebyshev nodes of order k on $I_{m,n}$.

We define $I_{m,n} = [2^{-m}n, 2^{-m}(n+1)]$ for $n = 0,1,\ldots,2^m - 1$, and obtain

$$\begin{aligned}
\|Q^k_m f - f\|^2 &= \int_0^1 \left[(Q^k_m f)(x) - f(x)\right]^2 dx \\
&= \sum_n \int_{I_{m,n}} \left[(Q^k_m f)(x) - f(x)\right]^2 dx \\
&\le \sum_n \int_{I_{m,n}} \left[(C^k_{m,n} f)(x) - f(x)\right]^2 dx \\
&\le \sum_n \int_{I_{m,n}} \left(\frac{2^{1-mk}}{4^k k!} \sup_{x\in I_{m,n}} |f^{(k)}(x)|\right)^2 dx \\
&\le \left(\frac{2^{1-mk}}{4^k k!} \sup_{x\in[0,1]} |f^{(k)}(x)|\right)^2,
\end{aligned}$$

and by taking square roots we have bound (7). Here $C^k_{m,n} f$ denotes the polynomial of degree k, which agrees with f at the Chebyshev nodes of order k on $I_{m,n}$, and we have used the well-known maximum error bound for Chebyshev interpolation (see, e.g., [5]). □

The error of the approximation $Q^k_m f$ of f, therefore, decays like 2^{-mk}, and, since S^k_m has a basis of $2^m k$ elements, we have convergence of order k. For the generalization to d dimensions, a similar argument shows that the rate of convergence is of order k/d.

2. Sparse representation of integral operators. The matrix representations of integral operators in multiwavelet bases are sparse (to finite precision) for the same class of integral operators as is treated in [4], namely, all Calderon–Zygmund and pseudodifferential operators. In applications, an operator kernel commonly has the form

$$K(x,t) = f(x,t)\, s(|x-t|) + g(x,t), \tag{8}$$

where f and g are analytic functions of x, t and s is analytic except at the origin where it is singular. In the following development we initially restrict ourselves to a simple example of this latter class of kernels, with $K(x,t) = \log|x-t|$. Although this kernel is symmetric and convolutional, neither of these properties is related to the sparsity. Instead, a proof of sparsity (presented in Lemma 2.2 below) relies solely on derivative estimates provided by the Cauchy integral formula for intervals separated from the singularity. Later we treat the more general situation of (8) with $s(x) = \log(x)$.

We begin this section by introducing some notation for integral equations.

2.1. Second-kind integral equations. A linear Fredholm integral equation of the second kind is an expression of the form

$$f(x) - \int_a^b K(x,t)\, f(t)\, dt = g(x), \tag{9}$$

where we assume that the kernel K is in $L^2[a,b]^2$ and the unknown f and right-hand side g are in $L^2[a,b]$. For notational simplicity, we restrict our attention to the interval $[a,b] = [0,1]$. We use the symbol $\mathcal{K}$ to denote the integral operator of (9), given by the formula

$$(\mathcal{K}f)(x) = \int_0^1 K(x,t)\, f(t)\, dt$$

for all $f \in L^2[0,1]$ and $x \in [0,1]$. Suppose that $\{b_1, b_2, \ldots\}$ is an orthonormal basis for $L^2[0,1]$; the expansion of K in this basis is given by the formula

$$K(x,t) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} K_{ij}\, b_i(x)\, b_j(t), \tag{10}$$

where the coefficient K_{ij} is given by the expression

$$K_{ij} = \int_0^1 \int_0^1 K(x,t)\, b_i(x)\, b_j(t)\, dx\, dt, \qquad i,j = 1,2,\ldots. \tag{11}$$

Similarly, the functions f and g have expansions

$$f(x) = \sum_{i=1}^{\infty} f_i\, b_i(x), \qquad g(x) = \sum_{i=1}^{\infty} g_i\, b_i(x),$$

where the coefficients f_i and g_i are given by the formulae

$$f_i = \int_0^1 f(x)\, b_i(x)\, dx, \qquad g_i = \int_0^1 g(x)\, b_i(x)\, dx, \qquad i = 1,2,\ldots.$$

The integral equation (9) then corresponds to the infinite system of equations

$$f_i - \sum_{j=1}^{\infty} K_{ij}\, f_j = g_i, \qquad i = 1,2,\ldots.$$

The expansion for K may be truncated at a finite number of terms, yielding the integral operator R defined by the formula

$$(Rf)(x) = \int_0^1 \sum_{i=1}^n \sum_{j=1}^n (K_{ij}\, b_i(x)\, b_j(t))\, f(t)\, dt, \qquad f \in L^2[0,1], \quad x \in [0,1],$$

which approximates $\mathcal{K}$. Integral equation (9) is thereby approximated by the system

$$f_i - \sum_{j=1}^n K_{ij}\, f_j = g_i, \qquad i = 1, \dots, n, \tag{12}$$

which is a system of n equations in n unknowns. The system (12) may be solved numerically to yield an approximate solution to (9), given by the expression

$$f_R(x) = \sum_{i=1}^n f_i\, b_i(x).$$

How large is the error $e_R = f - f_R$ of the approximate solution? We follow the derivation by Delves and Mohamed in [7]. Defining g_R by the formula

$$g_R(x) = \sum_{i=1}^n g_i\, b_i(x),$$

we rewrite (9) and (12) in terms of operators $\mathcal{K}$ and R to obtain

$$(I - \mathcal{K})f = g,$$
$$(I - R)f_R = g_R.$$

Combining the latter equations yields

$$(I - \mathcal{K})e_R = (\mathcal{K} - R)f_R + (g - g_R).$$

Provided that $(I - \mathcal{K})^{-1}$ exists, we obtain the error bound

$$\|e_R\| \le \|(I - \mathcal{K})^{-1}\| \cdot \|(\mathcal{K} - R)f_R + (g - g_R)\|. \tag{13}$$

The error depends, therefore, on the conditioning of the original integral equation, as is apparent from the term $\|(I - \mathcal{K})^{-1}\|$, on the fidelity of the finite-dimensional operator R to the integral operator $\mathcal{K}$, and on the approximation of g_R to g.

2.2. Representation in multiwavelet bases. We consider integral operators $\mathcal{K}$ with kernels that are analytic, except at $x = t$, where they are singular. In particular, we analyze singularities of the form $\log|x - t|$. An operator with such a kernel K, expanded in one of the multiwavelet bases defined above, is represented as a sparse matrix. This sparseness is due to the smoothness of K on rectangles separated from the "diagonal."

DEFINITION 2.1. We say that a rectangular region oriented parallel to the coordinate axes x, t is *separated from the diagonal* if its distance in the horizontal or vertical direction from the line $x = t$ is at least the length of its longer side. In symbols, a region $[x, x+a] \times [t, t+b] \subset \mathbf{R}^2$ is separated from the diagonal if $a + \max\{a, b\} \le t - x$ or $b + \max\{a, b\} \le x - t$. This definition is illustrated in Fig. 1.

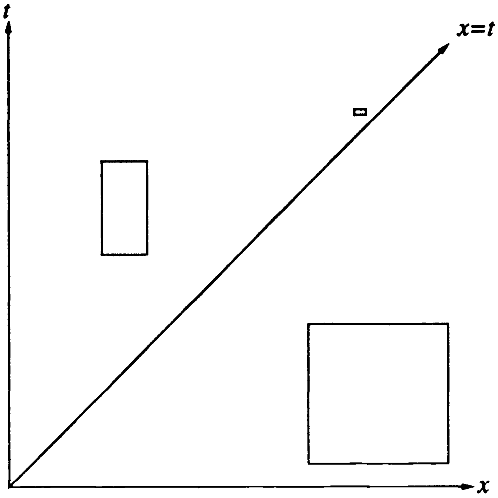


FIG. 1. *Rectangular regions (just) separated from the diagonal.*

Suppose that k is a positive integer and that $B_k = \{b_1, b_2, \ldots\}$ is the multiwavelet basis for $L^2[0,1]$ of order k, defined in §1. We let I_j denote the interval of support of b_j, and we assume that the sequence of basis functions $b_1, b_2, \ldots$ is ordered so that $I_1, I_2, \ldots$ have nonincreasing lengths. For large n, the matrix $\{K_{ij}\}_{i,j=1,\ldots,n}$ is sparse, to high precision, as is proved in the following propositions.

LEMMA 2.2. *Suppose that the function $K : [0,1] \times [0,1] \to \mathbf{R}$ is given by the formula $K(x,t) = \log|x-t|$. The expansion* (10) *of K in the multiwavelet basis B_k of order k has coefficients K_{ij} which satisfy the bound*

$$|K_{ij}| \le \frac{1}{8k \cdot 3^{k-1}} \tag{14}$$

whenever the rectangular region $I_i \times I_j$ is separated from the diagonal.

Proof. Suppose that the intervals I_i and I_j are given by the expressions $I_i = [x_0, x_0 + a]$ and $I_j = [t_0, t_0 + b]$; without loss of generality we assume (as one of two equivalent cases) that $b + \max\{a, b\} \le x_0 - t_0$. It is immediate from this inequality that

$$\left| \frac{x_0 + a/2 - x}{x_0 + a/2 - t} \right| \le \frac{1}{3} \tag{15}$$

for $(x,t) \in I_i \times I_j$.

We use the Taylor expansion for the natural logarithm about $c > 0$,

$$\log(c+y) = \log(c) + \left(\frac{y}{c}\right) - \frac{(y/c)^2}{2} + \frac{(y/c)^3}{3} - \frac{(y/c)^4}{4} + \cdots,$$

for $|y| < c$. We now let $c = x_0 + a/2 - t$ and $y = x - x_0 - a/2$ and for $(x,t) \in I_i \times I_j$ we obtain the formula

$$\log|x-t| = \log\left(x_0 + \left(\frac{a}{2}\right) - t\right) - \sum_{m=1}^{\infty} \frac{1}{m} \left(\frac{x_0 + a/2 - x}{x_0 + a/2 - t}\right)^m. \tag{16}$$

We now apply (11), (16), (2), and (15), each in turn, to obtain

$$\begin{aligned}
|K_{ij}| &= \left| \int_{t_0}^{t_0+b} \int_{x_0}^{x_0+a} K(x,t)\, b_i(x)\, b_j(t)\, dx\, dt \right| \\
&\le \int_{t_0}^{t_0+b} \left| \int_{x_0}^{x_0+a} \log|x-t|\, b_i(x)\, dx \right| |b_j(t)|\, dt \\
&= \int_{t_0}^{t_0+b} \left| \int_{x_0}^{x_0+a} \left[\log\left(x_0 + \frac{a}{2} - t\right) \right.\right. \\
&\qquad \left.\left. - \sum_{m=1}^{\infty} \frac{1}{m} \left(\frac{x_0 + a/2 - x}{x_0 + a/2 - t} \right)^m \right] b_i(x)\, dx \right| |b_j(t)|\, dt \\
&\le \int_{t_0}^{t_0+b} \left| \int_{x_0}^{x_0+a} \sum_{m=k}^{\infty} \frac{1}{m} \left(\frac{x_0 + a/2 - x}{x_0 + a/2 - t} \right)^m b_i(x)\, dx \right| |b_j(t)|\, dt \\
&\le \int_{t_0}^{t_0+b} \int_{x_0}^{x_0+a} \frac{1}{k} \sum_{m=k}^{\infty} \left(\frac{1}{3} \right)^m |b_i(x)|\, dx\, |b_j(t)|\, dt \\
&\le \int_{t_0}^{t_0+b} \int_{x_0}^{x_0+a} \frac{1}{2k \cdot 3^{k-1}} |b_i(x)|\, dx\, |b_j(t)|\, dt \\
&\le \frac{1}{2k \cdot 3^{k-1}} \int_{t_0}^{t_0+b} \sqrt{\left(\int_{x_0}^{x_0+a} {b_i}^2(x)\, dx \right) \left(\int_{x_0}^{x_0+a} 1\, dx \right)}\, |b_j(t)|\, dt \\
&\le \frac{\sqrt{ab}}{2k \cdot 3^{k-1}} \le \frac{1}{8k \cdot 3^{k-1}},
\end{aligned}$$

as was to be proved. □

We now consider a somewhat more general kernel.

LEMMA 2.3. *Suppose that the function $L : D \times D \to \mathbf{C}$ is given by the formula $L(z,w) = f(z,w) \log|z-w| + g(z,w)$, where D is the closed disk of radius $\frac{3}{2}$ centered at $z = \frac{1}{2}$ and f and g are analytic in a domain containing $D \times D \subset \mathbf{C}^2$. Suppose further that the function K is the restriction of L to $[0,1] \times [0,1]$. The expansion of K in the basis B_k has coefficients K_{ij} that satisfy the bound*

$$(17) \qquad |K_{ij}| \le \left(\frac{k}{8} + \frac{3}{16} \right) \frac{1}{3^{k-1}} \sup_{z,w \in \partial D} |f(z,w)| + \frac{2}{7 \cdot 8^k} \sup_{z,w \in \partial D} |g(z,w)|,$$

whenever the rectangular region $I_i \times I_j$ is separated from the diagonal.

Proof. We treat the parts of K separately by defining K' to be the restriction of $f(z,w) \log|z-w|$ to $[0,1] \times [0,1]$ and g' to be the restriction of g, so $K = K' + g'$.

We combine the method of proof used in Lemma 2.2 with the formula for the derivative of a product,

$$(18) \qquad \frac{\partial^m K'(x,t)}{\partial x^m} = \sum_{r=0}^{m} \binom{m}{r} \frac{\partial^r f(x,t)}{\partial x^r} \cdot \frac{\partial^{m-r} \log|x-t|}{\partial x^{m-r}}.$$

By the Cauchy integral formula we obtain

$$(19) \qquad \left| \frac{\partial^r f(x,t)}{\partial x^r} \right| \le r! \sup_{z,w \in \partial D} |f(z,w)|, \qquad \left| \frac{\partial^r g(x,t)}{\partial x^r} \right| \le r! \sup_{z,w \in \partial D} |g(z,w)|$$

for $(x,t) \in [0,1] \times [0,1]$. For the logarithm, differentiation yields the formula

$$\frac{\partial^{m-r} \log|x-t|}{\partial x^{m-r}} = \frac{(-1)^{m-r-1}(m-r-1)!}{(x-t)^{m-r}} \tag{20}$$

for $r < m$. Combining (18), (19), and (20), we obtain

$$\begin{aligned} \left|\frac{\partial^m K'(x,t)}{\partial x^m}\right| &\le \sum_{r=0}^{m} \binom{m}{r} \left|\frac{\partial^r f(x,t)}{\partial x^r}\right| \cdot \left|\frac{\partial^{m-r} \log|x-t|}{\partial x^{m-r}}\right| \\ &\le \sup_{z,w\in\partial D} |f(z,w)| \left(\sum_{r=0}^{m-1} \binom{m}{r} r! \frac{(m-r-1)!}{|x-t|^{m-r}} + m!\,|\log|x-t|| \right) \\ &\le S_f \cdot \left(m!\, \frac{2+\log m}{|x-t|^m} \right) \end{aligned} \tag{21}$$

for $|x-t| \le 1$ and $m \ge 1$, where $S_f = \sup_{z,w\in\partial D} |f(z,w)|$.

Suppose that the intervals I_i and I_j are given by the expressions $I_i = [x_0, x_0 + a]$ and $I_j = [t_0, t_0 + b]$; we assume without loss of generality that $b + \max\{a,b\} \le x_0 - t_0$. It follows directly from this inequality that

$$\left|\frac{x_0 + a/2 - x}{x_0 + a/2 - t}\right| \le \frac{1}{3} \tag{22}$$

for $(x,t) \in I_i \times I_j$. We now apply (11), (2), (21), and (22) to obtain

$$\begin{aligned} |K'_{ij}| &= \left| \int_{t_0}^{t_0+b} \int_{x_0}^{x_0+a} K'(x,t)\, b_i(x)\, b_j(t)\, dx\, dt \right| \\ &\le \int_{t_0}^{t_0+b} \left| \int_{x_0}^{x_0+a} \sum_{m=0}^{\infty} \frac{(x_0 + a/2 - x)^m}{m!} \frac{\partial^m K'(x_0 + a/2, t)}{\partial x_0{}^m}\, b_i(x)\, dx \right| |b_j(t)|\, dt \\ &\le \int_{t_0}^{t_0+b} \int_{x_0}^{x_0+a} \sum_{m=k}^{\infty} \left| \frac{x_0 + a/2 - x}{x_0 + a/2 - t} \right|^m S_f\, (2 + \log m)\, |b_i(x)|\, dx\, |b_j(t)|\, dt \\ &\le \int_{t_0}^{t_0+b} \int_{x_0}^{x_0+a} S_f \sum_{m=k}^{\infty} \left(\frac{1}{3}\right)^m (m+1)\, |b_i(x)|\, dx\, |b_j(t)|\, dt \\ &\le \int_{t_0}^{t_0+b} \int_{x_0}^{x_0+a} S_f \left(\frac{k}{2} + \frac{3}{4}\right) \frac{1}{3^{k-1}}\, |b_i(x)|\, dx\, |b_j(t)|\, dt \\ &\le S_f \left(\frac{k}{2} + \frac{3}{4}\right) \frac{1}{3^{k-1}} \int_{t_0}^{t_0+b} \sqrt{\left(\int_{x_0}^{x_0+a} b_i{}^2(x)\, dx \right) \left(\int_{x_0}^{x_0+a} 1\, dx \right)}\, |b_j(t)|\, dt \\ &\le S_f \left(\frac{k}{2} + \frac{3}{4}\right) \frac{\sqrt{ab}}{3^{k-1}} \\ &\le S_f \left(\frac{k}{8} + \frac{3}{16}\right) \frac{1}{3^{k-1}}. \end{aligned}$$

For the second term of $K_{ij} = K'_{ij} + g'_{ij}$ we obtain

$$\begin{aligned}
|g'_{ij}| &= \left| \int_{t_0}^{t_0+b} \int_{x_0}^{x_0+a} g'(x,t)\, b_i(x)\, b_j(t)\, dx\, dt \right| \\
&\le \int_{t_0}^{t_0+b} \left| \int_{x_0}^{x_0+a} g'(x,t)\, b_i(x)\, dx \right| |b_j(t)|\, dt \\
&\le \int_{t_0}^{t_0+b} \int_{x_0}^{x_0+a} \sum_{m=k}^{\infty} \sup_{z,w\in\partial D} |g(z,w)|\, |x - x_0 - a/2|^m\, |b_i(x)|\, dx |b_j(t)|\, dt \\
&\le \int_{t_0}^{t_0+b} \sum_{m=k}^{\infty} \sup_{z,w\in\partial D} |g(z,w)|\, \frac{1}{8^m} \int_{x_0}^{x_0+a} |b_i(x)|\, dx |b_j(t)|\, dt \\
&\le \frac{\sqrt{ab}}{7\cdot 8^{k-1}} \sup_{z,w\in\partial D} |g(z,w)| \\
&\le \frac{2}{7\cdot 8^{k}} \sup_{z,w\in\partial D} |g(z,w)|.
\end{aligned}$$

Combining the estimates for K'_{ij} and g'_{ij} yields (17). □

The preceding lemma shows that for a smooth kernel K with logarithm singularity at $x = t$, the order k of the multiwavelet basis B_k in which K is expanded may be chosen large enough that the expansion coefficient K_{ij} is negligible, provided $I_i \times I_j$ is separated from the diagonal. As mentioned above, a similar statement can be proven for any kernel of the form $K(x,t) = f(x,t)\, s(|x-t|) + g(x,t)$, where f, g are entire analytic functions of two variables and s is an analytic function except at the origin (where it has a singularity), provided that s is integrable. More generally, any Calderon–Zygmund or pseudodifferential operator can be similarly expressed (see [4]).

The next lemma establishes the fact that, asymptotically, most regions $I_i \times I_j$ *are* separated from the diagonal.

LEMMA 2.4. *Suppose that* $I_1, \ldots, I_n$ *are the* (*nonincreasing*) *intervals of support of the first* n *functions of the basis* B_k. *Of the* n^2 *rectangular regions* $I_i \times I_j$, *we denote the number separated from the diagonal by* $S(n)$ *and the number "near" the diagonal by* $N(n) = n^2 - S(n)$. *Then* $N(n)$ *grows as* $O(n \log n)$; *in particular, for* $n = 2^l k$ *with* $l > 0$, *we have the formula*

$$N(n) = 6nlk - 15nk - 6lk^2 + 16k^2. \tag{23}$$

Proof. The restriction that $n = 2^l k$ ensures that the first n basis functions consist of those functions whose intervals of support have length at least 2^{1-l}. We define $S_1(p)$ to be the number of pairs (i,j) such that the rectangular region $I_i \times I_j$ is separated from the diagonal and $|I_i| = |I_j| = 2^{-p}$, and we observe that $S_1(p) = (2^p - 1)(2^p - 2)k^2$ for $p = 0, 1, 2, \ldots$. We further define $S_2(p,q)$ to be the number of pairs (i,j) such that $I_i \times I_j$ is separated from the diagonal and $|I_i| = 2^{-p}$, $|I_j| = 2^{-q}$, and we observe that $S_2(p,q) = S_1(\min\{p,q\})\, 2^{|p-q|}$ for $p, q = 0, 1, 2, \ldots$. Finally, we combine these formulae to obtain

$$
\begin{aligned}
S(n) &= \sum_{p=0}^{l-1} \left(S_1(p) + \sum_{q=p+1}^{l-1} (S_2(p,q) + S_2(q,p)) \right) \\
&= \sum_{p=0}^{l-1} S_1(p)\ (1 + 2(2^{l-p} - 2)) \\
&= \sum_{p=0}^{l-1} (2^p - 1)(2^p - 2)\, k^2\, (2^{l-p+1} - 3) \\
&= (4^l - 6 \cdot 2^l l + 15 \cdot 2^l + 6l - 16)\, k^2 \\
&= n^2 - 6nlk + 15nk + 6lk^2 - 16k^2,
\end{aligned}
$$

from which (23) follows directly. The assertion that the general growth of $N(n)$ is $O(n \log n)$ follows from (23) and the fact that N is a monotonic function of n. □

3. Numerical examples and discussion.

3.1. Basis functions. In this section we give numerical expressions for the multiwavelet functions $f_0, f_1, \ldots, f_{k-1}$ and show their graphs for several values of k. These functions were obtained using the procedure of §1, implemented in a simple Maple program (available from the author). Table 1 contains, for small k, the polynomials that represent the f_i on the interval $(0,1)$, together with the reflection formula to extend the functions to $(-1,1)$, which is their interval of support. Figure 2 shows the graphs of the functions for $k = 4$ and $k = 5$.

3.2. Integral operators and their inverses. We compute the expansion in multiwavelet bases of the integral operator $\mathcal{K}$ defined by the formula

$$(\mathcal{K}f)(x) = \int_0^1 \log|x - t|\, f(t)\, dt, \tag{24}$$

which yields the matrix

$$K^{(n)} = \{K_{ij}\}_{i,j=1,\ldots,n},$$

where

$$K_{ij} = \int_0^1 \int_0^1 K(x,t)\, b_i(x)\, b_j(t)\, dx\, dt$$

and $\{b_1, b_2, \ldots\}$ are a multiwavelet basis of $L^2[0,1]$. We approximate $K^{(n)}$ with a matrix $T^{(n)}$ whose elements are defined by the formula

$$T_{ij}^{(n)} = \begin{cases} K_{ij}, & \text{if } |K_{ij}| \geq \tau, \\ 0, & \text{otherwise,} \end{cases} \tag{25}$$

where the threshold τ is chosen so that a desired precision ϵ is maintained: $\|T^{(n)} - K^{(n)}\| \leq \epsilon \|K^{(n)}\|$. Here the norm $\|\cdot\|$ is the row-sum norm, $\|A\| = \max_i \sum_{j=1}^n |A_{ij}|$. The threshold τ is given by $\tau = \epsilon \|K^{(n)}\|/n$. This computation was performed for the multiwavelet basis of order $k = 4$, for various sizes n, as shown in Table 2.

An interesting property of many operators of second-kind integral equations is that their inverses, when they exist, are also sparse in multiwavelet coordinates (to

TABLE 1

Expressions for the orthonormal, vanishing-moment functions $f_1, \ldots, f_k$, for various k, for argument x in the interval $(0,1)$. The function f_i is extended to the interval $(-1,1)$ as an odd or even function, according to the formula $f_i(x) = (-1)^{i+k-1} f_i(-x)$ for $x \in (-1,0)$, and is zero outside $(-1,1)$.

	$k=1$	
$f_1(x) =$	$\sqrt{\frac{1}{2}}$	
	$k=2$	
$f_1(x) =$	$\sqrt{\frac{3}{2}}$	$(-1+2x)$
$f_2(x) =$	$\sqrt{\frac{1}{2}}$	$(-2+3x)$
	$k=3$	
$f_1(x) =$	$\frac{1}{3}\sqrt{\frac{1}{2}}$	$(1-24x+30x^2)$
$f_2(x) =$	$\frac{1}{2}\sqrt{\frac{3}{2}}$	$(3-16x+15x^2)$
$f_3(x) =$	$\frac{1}{3}\sqrt{\frac{5}{2}}$	$(4-15x+12x^2)$
	$k=4$	
$f_1(x) =$	$\sqrt{\frac{15}{34}}$	$(1+4x-30x^2+28x^3)$
$f_2(x) =$	$\sqrt{\frac{1}{42}}$	$(-4+105x-300x^2+210x^3)$
$f_3(x) =$	$\frac{1}{2}\sqrt{\frac{35}{34}}$	$(-5+48x-105x^2+64x^3)$
$f_4(x) =$	$\frac{1}{2}\sqrt{\frac{5}{42}}$	$(-16+105x-192x^2+105x^3)$
	$k=5$	
$f_1(x) =$	$\sqrt{\frac{1}{186}}$	$(1+30x+210x^2-840x^3+630x^4)$
$f_2(x) =$	$\frac{1}{2}\sqrt{\frac{1}{38}}$	$(-5-144x+1155x^2-2240x^3+1260x^4)$
$f_3(x) =$	$\sqrt{\frac{35}{14694}}$	$(22-735x+3504x^2-5460x^3+2700x^4)$
$f_4(x) =$	$\frac{1}{8}\sqrt{\frac{21}{38}}$	$(35-512x+1890x^2-2560x^3+1155x^4)$
$f_5(x) =$	$\frac{1}{2}\sqrt{\frac{7}{158}}$	$(32-315x+960x^2-1155x^3+480x^4)$

high precision). The operator $(I-K)^{-1}$ has the Neumann expansion $\sum_{i=0}^{\infty} K^i$, which converges if $\|K\| < 1$; thus $(I-K)^{-1}$ may be approximated to arbitrary precision by a polynomial in K. More generally (regardless of $\|K\|$), $(I-K)^{-1} = A\sum_{i=0}^{\infty}(I-(I-K)A)^i$, where $A = (I-K^H)/\|(I-K^H)(I-K)\|$. The Schulz method [16] (see also [2]), a classical iterative matrix inversion technique, can be used to compute the first 2^m terms of this expansion with m iterations. Analogous to Newton iteration, the mth Schulz iterate X_m to invert a matrix M is given by $X_m = 2X_{m-1} - X_{m-1}\,M\,X_{m-1}$, where $X_0 = M^H/\|M^H M\|$. The iterates satisfy the equation $I - X_m M = (I - X_{m-1}M)^2$, which assures their quadratic convergence to M^{-1}.

The terms K^i and $((I-K^H)(I-K))^i$, of which these expansions are composed, have representations in multiwavelets that are asymptotically sparse. Specifically, their $n \times n$-matrix representations, after thresholding, contain only order $O(n \log n)$ nonzero elements. This fact follows from arguments similar to those given in Lemmas 2.2 and 2.3. It is important to add, however, that the constants in these asymptotic estimates may not ensure useful sparsity for reasonable values of n.

For the operator $T^{(n)}$ defined above, the inverse $(I-T^{(n)})^{-1}$ is roughly as sparse as $I-T^{(n)}$. We have computed it by the Schulz method. Table 2 displays, for various precisions ϵ, the average number of elements per row in the matrices $I-T^{(n)}$ and $(I-T^{(n)})^{-1}$. Figure 3 displays the matrices for $n = 128$ and $\epsilon = 10^{-3}$.

3.3. Discussion. The results of the previous subsection demonstrate, for a particular integral operator, that the multiwavelet representations are sparse. The matrix has a peculiar structure in which the nonnegligible elements are contained in blocks lying along rays emanating from one corner of the matrix. Furthermore, the inverse

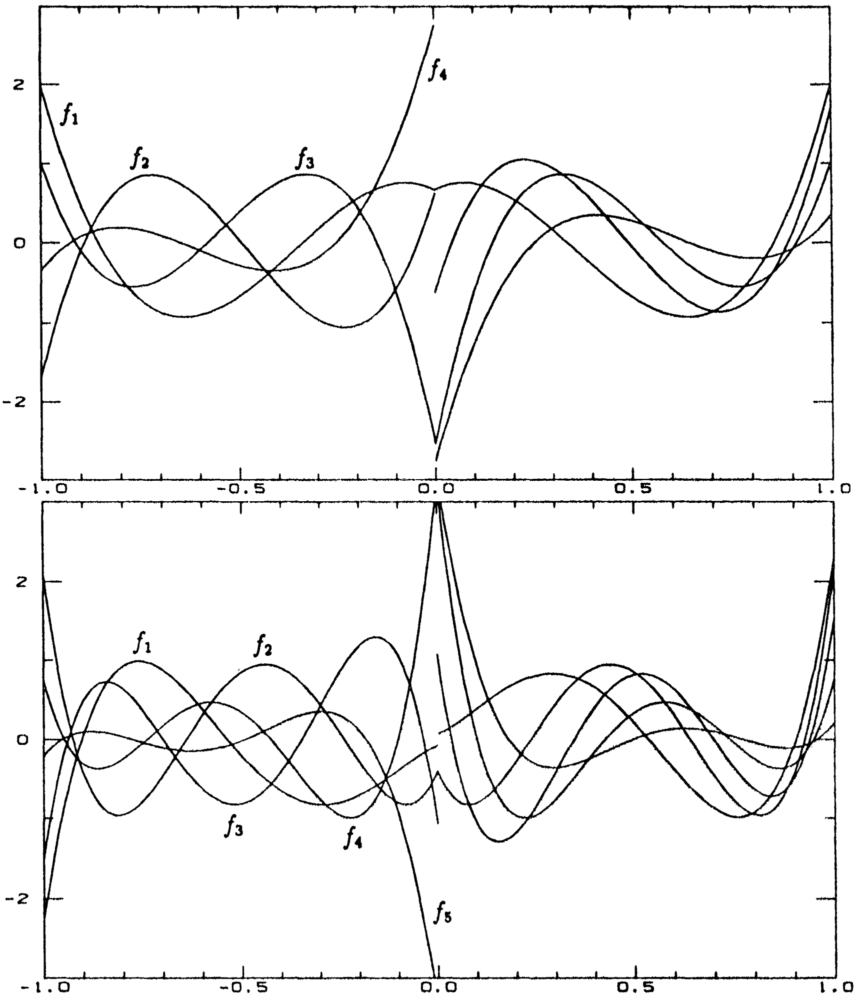


FIG. 2. *Functions $f_1, \ldots, f_k$ are graphed for $k = 4$ (top graph) and $k = 5$ (bottom). Each function (given in Table 1) is a polynomial on the interval $(0, 1)$, is an odd or even function on $(-1, 1)$, and is zero elsewhere.*

matrix shares that structure. This property is a general characteristic of integral operators with nonoscillatory kernels that possess diagonal singularities.

The kernel $K(x,t) = \log|x - t|$ of the previous subsection was chosen, however, because the projections K_{ij} could be computed analytically, thereby avoiding use of quadratures. The difficulty here with quadratures is that they would be required for each element K_{ij}, and would have to cope with the singularity of the logarithm. It was felt that the analytical computation would be more efficient. In fact, the analytical computation, which requires integrating monomials x^j $(0 \leq j < k)$ against the logarithm and combining the results with large coefficients, is a very poorly-conditioned procedure. The computations described above required quadruple-precision arithmetic to obtain single-precision accuracy for n as small as 64. This procedure is not recommended.

TABLE 2

The average number of elements per row of the matrices $S^{(n)} = I - T^{(n)}$ and $(S^{(n)})^{-1}$, where $T^{(n)}$ is defined in (25), is tabulated for various precisions ϵ and various sizes n. Here $k = 4$.

	$\epsilon = 10^{-2}$		$\epsilon = 10^{-3}$		$\epsilon = 10^{-4}$	
n	$S^{(n)}$	$(S^{(n)})^{-1}$	$S^{(n)}$	$(S^{(n)})^{-1}$	$S^{(n)}$	$(S^{(n)})^{-1}$
32	8.8	9.7	19.3	19.6	22.8	23.6
64	9.3	10.0	25.8	26.0	31.9	32.6
128	9.9	10.1	29.2	29.4	38.2	38.8
256	11.8	11.8	30.1	30.3	41.9	42.7

The fault lies, of course, not with the idea of projection to the multiwavelet basis, but with the method of projection. The integration should be performed numerically, with quadratures. As mentioned above, such a procedure would require use of quadratures for each matrix element K_{ij}, or potentially order $O(n \log n)$ times. A more efficient procedure is to use the Nyström method, in which only n quadrature applications are required. Numerical quadratures and a vector-space analogue of the multiwavelet bases are developed in [1], [2]; these tools enable efficient solution of second-kind integral equations using Nyström's method. We believe that the present paper, rather than directly providing numerical tools, offers a particularly simple framework in which to understand the ideas for sparse representation of integral operators.

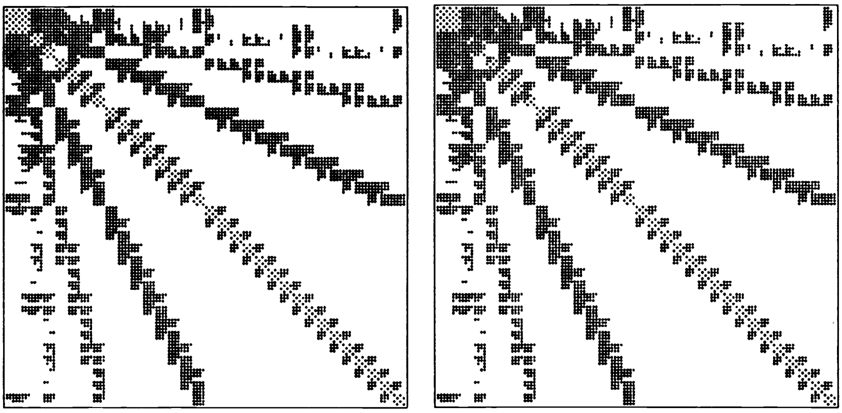


FIG. 3. *Matrices representing the operators $I - \mathcal{K}$ (left) and $(I - \mathcal{K})^{-1}$ (right), with $\mathcal{K}$ defined by (24), expanded in the multiwavelet basis of order $k = 4$, for $n = 128$. The dots represent elements above a threshold, which is determined so as to bound the relative truncation error at $\epsilon = 10^{-3}$.*

Acknowledgment. The bases constructed in this paper are the limiting case of the discrete construction in [2]; thanks to R. Coifman for prodding this author to consider the limit, which surprised us with its simplicity.

REFERENCES

[1] B. ALPERT, *Sparse Representation of Smooth Linear Operators*, Ph.D. thesis, Department of Computer Science, Yale University, New Haven, CT, 1990.

[2] B. ALPERT, G. BEYLKIN, R. COIFMAN, AND V. ROKHLIN, *Wavelets for the fast solution of second-kind integral equations*, Tech. Report, Dept. of Computer Science, Yale University, New Haven, CT, 1990; SIAM J. Sci. Comput., to appear.

[3] B. Alpert and V. Rokhlin, *A fast algorithm for the evaluation of Legendre expansions*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 158–179.

[4] G. Beylkin, R. Coifman, and V. Rokhlin, *Fast wavelet transforms and numerical algorithms* I, Comm. Pure Appl. Math., 44 (1991), pp. 141–183.

[5] G. Dahlquist and Å. Björck, *Numerical Methods*, Prentice Hall, Englewood Cliffs, NJ, 1974.

[6] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[7] L. M. Delves and J. L. Mohamed, *Computational Methods for Integral Equations*, Cambridge University Press, London, 1985.

[8] L. Greengard and V. Rokhlin, *A fast algorithm for particle simulations*, J. Comput. Phys., 73 (1987), pp. 325–348.

[9] A. Grossman and J. Morlet, *Decomposition of Hardy functions into square integrable wavelets of constant shape*, SIAM J. Math. Anal., 15 (1984), pp. 723–736.

[10] S. Mallat, *Multiresolution approximations and wavelet orthonormal bases of* $L^2(\mathbf{R})$, Trans. Amer. Math. Soc. 315 (1989), pp. 69–88.

[11] Y. Meyer, *Principe d'incertitude, bases Hilbertiennes et algèbres d'opérateurs*, Tech. Report, Séminaire Bourbaki, nr. 662, 1985–1986.

[12] ———, *Ondelettes et functions splines*, Tech. Report, Séminaire EDP, Ecole Polytechnique, Paris, 1986.

[13] Y. Meyer, Seminar presented at the Department of Mathematics, Yale University, New Haven, CT, June, 1990; Rev. Mat. Iberoamericana, to appear.

[14] S. O'Donnell and V. Rokhlin, *A fast algorithm for the numerical evaluation of conformal mappings*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 475–487.

[15] V. Rokhlin, *A fast algorithm for the discrete Laplace transformation*, J. Complexity, 4 (1988), pp. 12–32.

[16] G. Schulz, *Iterative berechnung der reziproken matrix*, Z. Angew. Math. Mech., 13 (1933), pp. 57–59.

SIAM J. MATH. ANAL.
Vol. 24, No. 1, pp. 263–277, January 1993

017

INEQUALITIES OF LITTLEWOOD–PALEY TYPE FOR FRAMES AND WAVELETS*

CHARLES K. CHUI[†] AND XIANLIANG SHI[†‡]

Abstract. Inequalities of Littlewood–Paley type for frames in both the wavelet and Weyl–Heisenberg settings, and those for any unconditional basis of the form $\psi_{j,k}(x) = 2^{\frac{j}{2}}\psi(2^j x - k)$, are established. In particular, if $\{\psi_{j,k}\}$ is a semi-orthogonal basis, then the Littlewood–Paley identity is obtained. A similar identity for the "biorthogonal wavelets" of Cohen, Daubechies, and Feauveau is also obtained.

Key words. Littlewood–Paley inequalities, frames, frame bounds, wavelets, biorthogonal wavelets

AMS(MOS) subject classifications. 41A17, 42C15

1. Introduction and results.

The objective of this paper is to establish certain inequalities and identities of Littlewood–Paley type and to discuss some of their important consequences. This section is devoted to introducing the necessary definitions and notation and to a discussion of the main results in this paper. For the sake of clarity, it will be divided into two subsections: with the first one on frames and the second one on wavelets.

1.1. Inequalities for frames.

The notion of frames was introduced by Duffin and Schaeffer [6] in their work on nonharmonic analysis. For the Hilbert space $L^2 := L^2(-\infty,\infty)$, a family of functions $\phi_k \in L^2$, $k \in \mathbb{Z}$, is said to be a frame of L^2 if there exist two positive constants, C_1 and C_2, with $0 < C_1 \le C_2 < \infty$, such that

$$C_1\|f\|^2 \le \sum_{k\in\mathbb{Z}} |\langle f, \phi_k\rangle|^2 \le C_2\|f\|^2 \tag{1.1}$$

for all $f \in L^2$. Here and throughout, $\|f\|$ denotes the L^2-norm of f. A frame $\{\phi_k\}$ is called a tight frame, if $C_1 = C_2$ (cf. [5], [6]). Note that even a tight frame with $C_1 = C_2 = 1$ is not necessarily a basis of L^2. For instance, if $\{\eta_k\}$, $k \in \mathbb{Z}$, is an orthonormal basis of L^2 and $\{\alpha_k\}$, $k \in \mathbb{Z}$, is any sequence of real numbers, then the family $\{\eta_k \cos\alpha_k, \eta_k \sin\alpha_k\}$, $k \in \mathbb{Z}$, which is certainly not a basis of L^2, is, however, a tight frame of L^2 with $C_1 = C_2 = 1$. Observe that $\{\alpha_k\}$ may be so chosen that every function in this nonbasis tight frame is nontrivial.

In this paper, we will only consider frames that are generated by a single function. Two types of such frames are of particular interest:

(i) *s.t. frames* (or frames generated by *scaling* and *translation* of a function $\psi \in L^2$) defined by:

$$(S_{j,k}\psi)(x) := a^{\frac{j}{2}}\psi(a^j x - kb), \qquad j,k \in \mathbb{Z}, \tag{1.2}$$

*Received by the editors May 28, 1991; accepted for publication (in revised form) April 21, 1992. This research was supported by National Science Foundation grants DMS-89-01345 and INT-87-12424.

[†] Center for Approximation Theory, Department of Mathematics, Texas A&M University, College Station, Texas 77843-3368.

[‡] Permanent address, Department of Mathematics, Hangzhou University, Hangzhou, Zhejiang, 310028, People's Republic of China.

where $a > 1$ and $b > 0$ are (fixed) constants, and

(ii) *w.h. frames* (or frames of Weyl–Heisenberg type, generated by a function $\phi \in L^2$) defined by

$$(H_{j,k}\phi)(x) := e^{ijpx}\phi(x - kq), \qquad j,k \in \mathbb{Z}, \tag{1.3}$$

where $p, q > 0$ and $pq \le 2\pi$.

1.1.1. s.t. frames. Let us first study s.t. frames. For any function $\psi \in L^2$, set

$$\psi_j(x) := a^j\overline{\psi(-a^jx)}, \qquad j \in \mathbb{Z}, \tag{1.4}$$

where again $a > 1$ is fixed. Then we have the so-called "semidiscrete integral wavelet transform":

$$(W_jf)(x) = (f * \psi_j)(x), \qquad j \in \mathbb{Z}, \quad f \in L^2, \tag{1.5}$$

where $*$ denotes the integral convolution on $(-\infty, \infty)$ (cf. [8], [9]). For this transform to have any practical value, it must be "stable"; and by this, we mean the existence of constants A and B, with $0 < A \le B < \infty$, such that

$$A\|f\|^2 \le \sum_{j\in\mathbb{Z}} \|W_jf\|^2 \le B\|f\|^2, \qquad f \in L^2. \tag{1.6}$$

On the other hand, in signal analysis, since ψ has the property of a bandpass filter, in order to be able to reconstruct the original signal from its wavelet transform (1.5), the function ψ must satisfy

$$A \le \sum_{j\in\mathbb{Z}} |\widehat{\psi}(a^j\omega)|^2 \le B \quad \text{a.e.}, \tag{1.7}$$

where $\widehat{\psi}$ is the Fourier transform of ψ, defined by

$$\widehat{\psi}(\omega) = \int_{-\infty}^{\infty} e^{-ix\omega}\psi(x)dx.$$

In fact, it is easy to see that (1.6) and (1.7) are equivalent, with the same constants A and B.

To generate an s.t. frame, we further discretize the transform W_jf in (1.5) by defining $S_{j,k}\psi$ as in (1.2), using another parameter $b > 0$. Then analogous to the stability condition (1.6) for semidiscrete integral wavelet transforms, we require

$$\psi_{j,k}(x) := (S_{j,k}\psi)(x) = a^{\frac{j}{2}}\psi(a^jx - kb) \tag{1.8}$$

to satisfy the frame condition:

$$A'\|f\|^2 \le \sum_{j,k\in\mathbb{Z}} |\langle f, \psi_{j,k}\rangle|^2 \le B'\|f\|^2, \qquad f \in L^2, \tag{1.9}$$

where $0 < A' \le B' < \infty$. Our result on s.t. frames is then an analogue of (1.7), as follows.

THEOREM 1. *Let $\{\psi_{j,k}\}$, as defined in* (1.8) *for some $a > 1$ and $b > 0$, be a frame of L^2 with frame bounds A' and B' as in* (1.9). *Then $\widehat{\psi}$ satisfies*

$$A' \leq \frac{1}{b}\sum_{j\in\mathbb{Z}}|\widehat{\psi}(a^j\omega)|^2 \leq B' \quad a.e., \tag{1.10}$$

for the same constants A' and B'.

From (1.10), we can easily derive other interesting inequalities. For instance, by integrating each term in

$$\frac{A'}{|\omega|} \leq \frac{1}{b}\sum_{j\in\mathbb{Z}}\frac{|\widehat{\psi}(a^j\omega)|^2}{|\omega|} \leq \frac{B'}{|\omega|}$$

over $1 \leq |\omega| \leq a$, we have

$$2A'\log a \leq \frac{1}{b}\sum_{j\in\mathbb{Z}}\int_{1\leq|\omega|\leq a}\frac{|\widehat{\psi}(a^j\omega)|^2}{|\omega|}d\omega \leq 2B'\log a,$$

which immediately yields

$$A' \leq \frac{1}{2b\log a}\int_{-\infty}^{\infty}\frac{|\widehat{\psi}(\omega)|^2}{|\omega|}d\omega \leq B'. \tag{1.11}$$

We remark that the so-called "compactibility condition" (1.11) for s.t. frames was also derived by Daubechies [5] by using techniques from trace-class operators. In addition, Daubechies [5] also observed that under the assumptions of Theorem 1, the quantity

$$\sum_{j\in\mathbb{Z}}|\widehat{\psi}(a^j\omega)|^2$$

is bounded from above and below by some positive constants. The contribution in Theorem 1 is that these constants are given by the frame bounds.

1.1.2. w.h. frames. For a function $\phi \in L^2$, consider the semidiscrete window Fourier transform (also known as short-time Fourier transform):

$$(F_jf)(x) = \int_{-\infty}^{\infty}e^{-ijpt}\overline{\phi(t-x)}f(t)dt, \qquad f \in L^2, \tag{1.12}$$

where $p > 0$ is a fixed constant. Then analogous to (1.6), the stability of the transform (1.12) is defined by the requirement:

$$C\|f\|^2 \leq \sum_{j\in\mathbb{Z}}\|F_jf\|^2 \leq D\|f\|^2, \qquad f \in L^2, \tag{1.13}$$

for some $0 < C \leq D < \infty$, independent of f. It is not difficult to see that (1.13) is equivalent to

$$C \leq \sum_{j\in\mathbb{Z}}|\hat{\phi}(\omega - jp)|^2 \leq D \quad \text{a.e.}, \tag{1.14}$$

for the same constants C and D. Further discretization of the transform $F_j f$ in (1.12) results in introducing

$$\phi^{j,k}(x) := (H_{j,k}\phi)(x) = e^{ijpx}\phi(x - kq), \tag{1.15}$$

as defined by (1.3), where $q > 0$ is the second discretization constant satisfying

$$0 < pq \leq 2\pi \tag{1.16}$$

(cf. [5] for the requirement of (1.16)). For $\{\phi^{j,k}\}$ to be a frame, we need constants $0 < C' \leq D' < \infty$, such that

$$C'\|f\|^2 \leq \sum_{j,k\in\mathbb{Z}} |\langle f, \phi^{j,k}\rangle|^2 \leq D'\|f\|^2, \qquad f \in L^2. \tag{1.17}$$

Our result on w.h. frames in this paper is the following inequalities, which are along the same line of (1.14).

THEOREM 2. *Let* $\{\phi^{j,k}\}$*, as defined in* (1.15) *for some* $p, q > 0$ *satisfying* (1.16)*, be a frame of* L^2 *with frame bounds* C' *and* D' *as in* (1.17)*. Then* ϕ *and* $\hat{\phi}$ *satisfy*

$$C' \leq \frac{1}{q}\sum_{j\in\mathbb{Z}} |\hat{\phi}(\omega - jp)|^2 \leq D', \quad \textit{and} \tag{1.18}$$

$$C' \leq \frac{2\pi}{p}\sum_{j\in\mathbb{Z}} |\phi(x - jq)|^2 \leq D' \quad \textit{a.e.}, \tag{1.19}$$

for the same constants C' *and* D'*.*

1.2. Results on wavelets. In the following, we will set $a = 2$ and $b = 1$ in the definition of $\psi_{j,k}$ in (1.8); that is, we will consider

$$\psi_{j,k}(x) = 2^{\frac{j}{2}}\psi(2^j x - k), \qquad j, k \in \mathbb{Z}. \tag{1.20}$$

Furthermore, we will also assume that $\{\psi_{j,k}\}$ in (1.20) is an unconditional basis of L^2 with constants $0 < K \leq L < \infty$, namely: it is complete and satisfies

$$K \sum_{j,k\in\mathbb{Z}} |a_{j,k}|^2 \leq \left\| \sum_{j,k\in\mathbb{Z}} a_{j,k}\psi_{j,k} \right\|^2 \leq L \sum_{j,k\in\mathbb{Z}} |a_{j,k}|^2 \tag{1.21}$$

for all $\{a_{j,k}\} \in \ell^2(\mathbb{Z}^2)$. However, this assumption does not guarantee that the dual basis $\{\psi^*_{j,k}\}$, relative to $\{\psi_{j,k}\}$, defined by

$$\psi^*_{j,k} \in L^2 \quad \text{and} \quad \langle \psi_{j,k}, \psi^*_{\ell,m}\rangle = \delta_{j,\ell}\delta_{k,m}, \tag{1.22}$$

is obtained by dyadic dilations and integral translations of a single function in the same manner as $\{\psi_{j,k}\}$ from ψ. We will give an elementary proof of this somewhat surprising result in §3. If it so happens that

$$\psi^*_{j,k} = \widetilde{\psi}_{j,k}, \qquad j, k \in \mathbb{Z}, \tag{1.23}$$

where

$$\widetilde{\psi}_{j,k}(x) := 2^{\frac{j}{2}}\widetilde{\psi}(2^j x - k) \tag{1.24}$$

for some $\widetilde{\psi} \in L^2$, then we will call $\widetilde{\psi}$ the "dual" of ψ. Since $\{\widetilde{\psi}_{j,k}\}$ is clearly an unconditional basis of L^2, it follows that ψ is the dual of $\widetilde{\psi}$, also. Observe that if ψ is an "orthonormal wavelet" in the sense that $\{\psi_{j,k}\}$ is an orthonormal basis of L^2, then ψ is self-dual with $\widetilde{\psi} = \psi$.

As a consequence of Theorem 1, we have the following inequalities of Littlewood–Paley type.

COROLLARY 3. *Let $\{\psi_{j,k}\}$, as defined in (1.20), be an unconditional basis of L^2, with bounds K and L as given in (1.21). Then the Fourier transform $\widehat{\psi}$ of ψ satisfies*

$$K \le \sum_{j\in\mathbb{Z}} |\widehat{\psi}(2^j\omega)|^2 \le L \quad a.e., \tag{1.25}$$

and consequently,

$$K \le \frac{1}{2\log 2}\int_{-\infty}^{\infty} \frac{|\widehat{\psi}(\omega)|^2}{|\omega|}d\omega \le L. \tag{1.26}$$

Furthermore, if ψ is a wavelet with dual $\widetilde{\psi}$ as defined in (1.22)–(1.24), then $\{\widetilde{\psi}_{j,k}\}$ is also an unconditional basis of L^2 with bounds L^{-1} and K^{-1}, and consequently,

$$L^{-1} \le \sum_{j\in\mathbb{Z}} |\widehat{\widetilde{\psi}}(2^j\omega)|^2 \le K^{-1} \quad a.e., \tag{1.27}$$

and

$$L^{-1} \le \frac{1}{2\log 2}\int_{-\infty}^{\infty} \frac{|\widehat{\widetilde{\psi}}(\omega)|^2}{|\omega|}d\omega \le K^{-1}. \tag{1.28}$$

There are two special cases that are of particular importance. We will discuss them separately.

1.2.1. Semi-orthogonal wavelets. Let $\{\psi_{j,k}\}$ be an unconditional basis of L^2 generated by some function ψ, as governed by (1.20) and (1.21). For each $j \in \mathbb{Z}$, set

$$W_j = \operatorname{clos}_{L^2} \operatorname{span}\{\psi_{j,k}\colon\ k \in \mathbb{Z}\}. \tag{1.29}$$

Then we say that ψ is a *"semi-orthogonal wavelet"* if

$$W_j \perp W_\ell, \qquad j,\ell \in \mathbb{Z}, \quad j \neq \ell. \tag{1.30}$$

The dual $\widetilde{\psi}$ of a semi-orthogonal wavelet ψ is easily obtained via the Fourier transform, namely

$$\widehat{\widetilde{\psi}}(\omega) = \frac{\widehat{\psi}(\omega)}{\Psi(\omega)}, \tag{1.31}$$

where

$$\Psi(\omega) := \sum_{j\in\mathbb{Z}} |\widehat{\psi}(\omega + 2\pi j)|^2. \tag{1.32}$$

By giving up the orthogonality of $\{\psi_{0,k}\colon\ k \in \mathbb{Z}\}$ it is possible to construct compactly supported ψ with certain desirable properties. For instance, the compactly supported spline-wavelets of Chui and Wang in [1] are symmetric for splines of even order and anti-symmetric for splines of order. In addition, explicit formulas of compactly supported semi-orthogonal spline wavelets and their duals were given in Chui and Wang [1], and a characterization of all the compactly supported ones in [2]. For this type of wavelets, we have the following result.

THEOREM 4. *Let ψ be a semi-orthogonal wavelet with dual $\tilde{\psi}$. Then*

$$\sum_{j\in\mathbb{Z}} \overline{\widehat{\psi}(2^j\omega)}\ \widehat{\tilde{\psi}}(2^j\omega) = 1 \quad a.e., \tag{1.33}$$

and consequently,

$$\int_{-\infty}^{\infty} \frac{\overline{\widehat{\psi}(\omega)}\ \widehat{\tilde{\psi}}(\omega)}{|\omega|} d\omega = 2\log 2. \tag{1.34}$$

In particular, if ψ is an orthonormal wavelet so that $\tilde{\psi} = \psi$, then its Fourier transform satisfies

$$\sum_{j\in\mathbb{Z}} |\widehat{\psi}(2^j\omega)|^2 = 1 \quad \text{a.e.}, \tag{1.35}$$

and

$$C_\psi := \int_{-\infty}^{\infty} \frac{|\widehat{\psi}(\omega)|^2}{|\omega|} d\omega = 2\log 2. \tag{1.36}$$

1.2.2. Nonorthogonal wavelets. Wavelets without any orthogonality structure have even more flexibility. For instance, examples of those ψ and their duals $\tilde{\psi}$ in $C^n(-\infty,\infty)$, where $n \in \mathbb{Z}_+$ is arbitrary, both with symmetry and compact support, have been constructed recently by Cohen, Daubechies, and Feauveau [4]. Following Cohen [3], we let $m_0(\omega)$ and $\tilde{m}_0(\omega)$ be two 2π-periodic Lip(α) continuous functions, $0 < \alpha < 1$, satisfying

$$\begin{gathered} \overline{m_0(\omega)}\tilde{m}_0(\omega) + \overline{m_0(\omega+\pi)}\tilde{m}_0(\omega+\pi) = 1; \\ m_0(0) = \tilde{m}_0(0) = 1; \qquad m_0(\pi) = \tilde{m}_0(\pi) = 1. \end{gathered} \tag{1.37}$$

Suppose that ϕ and $\tilde{\phi}$, defined by

$$\begin{aligned} \hat{\phi}(\omega) &= \prod_{j=1}^{\infty} m_0\left(\frac{\omega}{2^j}\right), \\ \hat{\tilde{\phi}}(\omega) &= \prod_{j=1}^{\infty} \tilde{m}_0\left(\frac{\omega}{2^j}\right), \end{aligned} \tag{1.38}$$

satisfy

$$|\phi(x)| + |\tilde{\phi}(x)| \le \frac{C}{1+|x|^{1+\varepsilon}}, \qquad x \in (-\infty,\infty),$$

for some $\varepsilon, C > 0$. Then ψ is a wavelet with dual $\widetilde{\psi}$, where

$$\begin{cases} \widehat{\psi}(\omega) = e^{-i\frac{\omega}{2}}\overline{\widetilde{m}_0\left(\frac{\omega}{2}+\pi\right)}\hat{\phi}\left(\frac{\omega}{2}\right), \\ \widehat{\widetilde{\psi}}(\omega) = e^{-i\frac{\omega}{2}}\overline{m_0\left(\frac{\omega}{2}+\pi\right)}\hat{\tilde{\phi}}\left(\frac{\omega}{2}\right), \end{cases} \tag{1.39}$$

in the sense that $\{\psi_{j,k}\}$ and $\{\widetilde{\psi}_{j,k}\}$, defined by (1.20) and (1.24), are both unconditional bases of L^2, such that

$$\langle \psi_{j,k}, \widetilde{\psi}_{\ell,m}\rangle = \delta_{j,\ell}\delta_{k,m},$$

(cf. Cohen [3]). Although this ψ is not semi-orthogonal, its Fourier transform still satisfies the same Littlewood–Paley identity, as follows.

THEOREM 5. *Let ψ be a wavelet with dual $\widetilde{\psi}$ as defined by* (1.37)–(1.39). *Then*

$$\sum_{j\in\mathbb{Z}} \overline{\widehat{\psi}(2^j\omega)}\;\widehat{\widetilde{\psi}}(2^j\omega) = 1, \qquad \omega \in \mathbf{R}, \tag{1.40}$$

and consequently,

$$\int_{-\infty}^{\infty} \frac{\overline{\widehat{\psi}(\omega)}\;\widehat{\widetilde{\psi}}(\omega)}{|\omega|} d\omega = 2\log 2. \tag{1.41}$$

Remark. For an orthonormal wavelet ψ, identity (1.35) is called the Littlewood–Paley identity. Hence, the results in (1.33) and (1.40) may be called Littlewood–Paley identities also. The significance of these identities is that when $\widetilde{\psi}$ is considered as a bandpass filter, so that $\{\widetilde{\psi}_{j,k}\}$ gives rise to a filter bank decomposition, the wavelet ψ can be used for perfect reconstruction. In other words, the pair $(\psi, \widetilde{\psi})$ constitutes an allpass filter as demonstrated by (1.33) and (1.40). We also remark that the value of C_ψ in (1.36) is needed in the reconstruction formula (from the integral wavelet transform) of Grossmann and Morlet.

2. Proofs.

In this section, we establish all the results stated in §1.

2.1. Proof of Theorem 1.

By (1.2), we have, for any $f \in L^2$,

$$\begin{aligned} \langle f, \psi_{j,k}\rangle &= a^{\frac{j}{2}} \int_{-\infty}^{\infty} f(x)\overline{\psi(a^j x - kb)}dx \\ &= \frac{1}{2\pi} a^{\frac{j}{2}} \int_{-\infty}^{\infty} \hat{f}(a^j\omega)\overline{\widehat{\psi}(\omega)}e^{ikb\omega}d\omega. \end{aligned}$$

Hence, by setting

$$T := \frac{2\pi}{b}, \tag{2.1}$$

we have

$$\begin{aligned}\sum_{j,k\in\mathbb{Z}}|\langle f,\psi_{j,k}\rangle|^2 &= \sum_{j\in\mathbb{Z}}\frac{a^j}{4\pi^2}\sum_{k\in\mathbb{Z}}\left|\int_{-\infty}^{\infty}\hat{f}(a^j\omega)\overline{\hat{\psi}(\omega)}e^{ikb\omega}d\omega\right|^2\\
&= \sum_{j\in\mathbb{Z}}\frac{a^jT^2}{4\pi^2}\sum_{k\in\mathbb{Z}}\left|\frac{1}{T}\int_0^T\left[\overline{\sum_{\ell\in\mathbb{Z}}\hat{f}(a^j(\omega+\ell T))\hat{\psi}(\omega+\ell T)}\right]e^{ik\frac{2\pi}{T}\omega}d\omega\right|^2 \qquad (2.2)\\
&= \sum_{j\in\mathbb{Z}}\frac{a^j}{2\pi b}\int_0^T\left|\sum_{\ell\in\mathbb{Z}}\hat{f}(a^j(\omega+\ell T))\overline{\hat{\psi}(\omega+\ell T)}\right|^2 d\omega.\end{aligned}$$

Therefore, it follows from (1.9) and (2.2) that

$$A'\|\hat{f}\|^2 \le \sum_{j\in\mathbb{Z}}\frac{a^j}{b}\int_0^T\left|\sum_{\ell\in\mathbb{Z}}\hat{f}(a^j(\omega+\ell T))\overline{\hat{\psi}(\omega+\ell T)}\right|^2 d\omega \le B'\|\hat{f}\|^2, \tag{2.3}$$

so that for any $M > 0$, $M \in Z$, and $\omega_0 \in (-\infty,\infty)$, we have

$$\sum_{j=-M}^{M}\frac{a^j}{b}\int_{a^{-j}\omega_0-\frac{T}{2}}^{a^{-j}\omega_0+\frac{T}{2}}\left|\sum_{\ell\in\mathbb{Z}}\hat{f}(a^j(\omega+\ell T))\overline{\hat{\psi}(\omega+\ell T)}\right|^2 d\omega \le B'\|\hat{f}\|^2.$$

Now, consider $\hat{f} = (1/\sqrt{2\varepsilon})\chi_{[\omega_0-\varepsilon,\omega_0+\varepsilon]}$, $\varepsilon > 0$. Then for sufficiently small ε, the above inequality becomes

$$\sum_{j=-M}^{M}\frac{a^j}{2\varepsilon b}\int_{a^{-j}(\omega_0-\varepsilon)}^{a^{-j}(\omega_0+\varepsilon)}|\hat{\psi}(\omega)|^2 d\omega \le B',$$

and thus, by taking $\varepsilon \to 0$ and $M \to \infty$ consecutively, we have

$$\frac{1}{b}\sum_{j\in\mathbb{Z}}|\hat{\psi}(a^j\omega)|^2 \le B' \quad \text{a.e.} \tag{2.4}$$

On the other hand, for any ω_0, $\eta > 0$, a positive integer M may be chosen so that

$$\int_{2a^M\omega_0(1+a)^{-1}}^{\infty}|\hat{\psi}(\omega)|^2 d\omega < \eta. \tag{2.5}$$

Also, for

$$0 < \varepsilon < \min\left\{\frac{a-1}{a+1}\omega_0, \frac{T}{2}\right\},$$

the function $\hat{f} = (1/\sqrt{2\varepsilon})\chi_{[\omega_0-\varepsilon,\omega_0+\varepsilon]}$ satisfies

$$\hat{f}(a^j(\omega+\ell T)) = 0$$

for all $\ell \in \mathbb{Z}$ with $|\ell| \geq (\varepsilon/a^j T)+1$ and all $\omega \in [a^{-j}\omega_0-(T/2), a^{-j}\omega_0+(T/2)]$. Hence, for this $\hat{f}$, we have

$$\begin{aligned}
&\sum_{j=-\infty}^{-M} \frac{a^j}{b} \int_{a^{-j}\omega_0-\frac{T}{2}}^{a^{-j}\omega_0+\frac{T}{2}} \left| \sum_{\ell\in\mathbb{Z}} \hat{f}(a^j(\omega+\ell T))\overline{\widehat{\psi}(\omega+\ell T)} \right|^2 d\omega \\
&\leq \sum_{j=-\infty}^{-M} \frac{a^j}{2\varepsilon b} \int_{a^{-j}\omega_0-\frac{T}{2}}^{a^{-j}\omega_0+\frac{T}{2}} \\
&\qquad \left[\sum_{\ell\in\mathbb{Z}} |\widehat{\psi}(\omega+\ell T)|^2 \chi_{[\omega_0-\varepsilon,\omega_0+\varepsilon]}(a^j(\omega+\ell T)) \right] \left(\frac{\varepsilon}{a^j T}+1 \right) d\omega \\
&\leq C \sum_{j=-\infty}^{-M} \int_{a^{-j}(\omega_0-\varepsilon)}^{a^{-j}(\omega_0+\varepsilon)} \left\{ |\widehat{\psi}(\omega)|^2 + \frac{a^j}{2\varepsilon}|\widehat{\psi}(\omega)|^2 \right\} d\omega.
\end{aligned} \tag{2.6}$$

Since $\varepsilon < (a-1/a+1)\omega_0$, the intervals

$$[a^{-j}(\omega_0-\varepsilon), a^{-j}(\omega_0+\varepsilon)], \qquad j \in \mathbb{Z},$$

are mutually disjoint; and hence, by (2.5), we have

$$\sum_{j=-\infty}^{-M} \int_{a^{-j}(\omega_0-\varepsilon)}^{a^{-j}(\omega_0+\varepsilon)} |\widehat{\psi}(\omega)|^2 d\omega \leq \int_{2a^M\omega_0(1+a)^{-1}}^{\infty} |\widehat{\psi}(\omega)|^2 d\omega < \eta,$$

so that it follows from (2.6) that

$$\begin{aligned}
&\sum_{j\in\mathbb{Z}} \frac{a^j}{b} \int_0^T \left| \sum_{\ell\in\mathbb{Z}} \hat{f}(a^j(\omega+\ell T))\overline{\widehat{\psi}(\omega+\ell T)} \right|^2 d\omega \\
&\leq \sum_{j=-M+1}^{\infty} \frac{a^j}{b} \int_0^T \left| \sum_{\ell\in\mathbb{Z}} \hat{f}(a^j(\omega+\ell T))\overline{\widehat{\psi}(\omega+\ell T)} \right|^2 d\omega \\
&\qquad + C\eta + \frac{C}{2\varepsilon} \int_{\omega_0-\varepsilon}^{\omega_0+\varepsilon} \sum_{j=-\infty}^{-M} |\widehat{\psi}(a^{-j}\omega)|^2 d\omega.
\end{aligned} \tag{2.7}$$

Therefore, by (2.3) and (2.7), we have

$$\begin{aligned}
I &:= \sum_{j=-M+1}^{\infty} \frac{a^j}{b} \int_0^T \left| \sum_{\ell\in\mathbb{Z}} \hat{f}(a^j(\omega+\ell T))\overline{\widehat{\psi}(\omega+\ell T)} \right|^2 d\omega \\
&\geq A' - C\eta - \frac{C}{2\varepsilon} \int_{\omega_0-\varepsilon}^{\omega_0+\varepsilon} \sum_{j=-\infty}^{-M} |\widehat{\psi}(a^{-j}\omega)|^2 d\omega.
\end{aligned} \tag{2.8}$$

On the other hand, for all sufficiently small $\varepsilon > 0$, it is clear that

$$\begin{aligned}
I &= \sum_{j=-M+1}^{\infty} \frac{a^j}{b} \int_{a^{-j}(\omega_0-\varepsilon)}^{a^{-j}(\omega_0+\varepsilon)} |\hat{f}(a^j\omega)\overline{\widehat{\psi}(\omega)}|^2 d\omega \\
&= \frac{1}{2b\varepsilon} \int_{\omega_0-\varepsilon}^{\omega_0+\varepsilon} \sum_{j=-M+1}^{\infty} |\widehat{\psi}(a^{-j}\omega)|^2 d\omega,
\end{aligned}$$

where $\hat{f} = (1/\sqrt{2\varepsilon})\chi_{[\omega_0-\varepsilon,\omega_0+\varepsilon]}$. Hence, in view of the boundedness property in (2.4), we may take $\varepsilon \to 0$ in (2.8) to arrive at

$$\sum_{j=-M+1}^{\infty} \frac{1}{b}|\hat{\psi}(a^{-j}\omega_0)|^2 \geq A' - C\eta - C\sum_{j=-\infty}^{-M} |\hat{\psi}(a^{-j}\omega_0)|^2 \tag{2.9}$$

for almost all $\omega_0 > 0$. Since $\eta > 0$ is arbitrary, (2.4) and (2.9) together yield

$$\frac{1}{b}\sum_{j\in\mathbb{Z}} |\hat{\psi}(a^{-j}\omega)|^2 \geq A' \tag{2.10}$$

for almost all $\omega > 0$. A similar argument holds for $\omega < 0$. Hence, by (2.4) and (2.10), we have completed the proof of Theorem 1.

2.2. Proof of Theorem 2. Instead of (2.1), we now consider $T = 2\pi/q$. Then it follows from (1.15) that, for any $f \in L^2$,

$$\begin{aligned}\sum_{j,k\in\mathbb{Z}} |\langle f, \phi^{j,k}\rangle|^2 &= \sum_{j,k\in\mathbb{Z}} \frac{1}{4\pi^2}\left|\int_{-\infty}^{\infty} e^{iq\omega(k-jp)}\overline{\hat{\phi}(\omega - jp)}\hat{f}(\omega)d\omega\right|^2 \\ &= \frac{T^2}{4\pi^2}\sum_{j,k\in\mathbb{Z}}\left|\frac{1}{T}\int_0^T \sum_{\ell\in\mathbb{Z}}\left\{\overline{\hat{\phi}(\omega+\ell T - jp)}\hat{f}(\omega+\ell T)\right\} e^{ik\frac{2\pi}{T}\omega}d\omega\right|^2 \\ &= \frac{T^2}{4\pi^2}\sum_{j\in\mathbb{Z}}\frac{1}{T}\int_0^T \left|\sum_{\ell\in\mathbb{Z}} \overline{\hat{\phi}(\omega+\ell T - jp)}\hat{f}(\omega+\ell T)\right|^2 d\omega.\end{aligned} \tag{2.11}$$

As in the proof of Theorem 1, we fix any ω_0 and $0 < \varepsilon < T/2$, and consider the function $\hat{f} = (1/\sqrt{2\varepsilon})\chi_{[\omega_0-\varepsilon,\omega_0+\varepsilon]}$. Then it follows from the frame bounds in (1.17) that (2.11) becomes

$$C' \leq \frac{1}{2\varepsilon}\int_{\omega_0-\varepsilon}^{\omega_0+\varepsilon} \sum_{j\in\mathbb{Z}} \frac{1}{q}|\hat{\phi}(\omega - jp)|^2 d\omega \leq D',$$

and this implies (1.18).

To derive (1.19), we set $T' = 2\pi/p$ and note that

$$\begin{aligned}\sum_{j,k\in\mathbb{Z}} |\langle f, \phi^{j,k}\rangle|^2 &= T'^2 \sum_{j,k\in\mathbb{Z}}\left|\frac{1}{T'}\int_0^{T'} \sum_{\ell\in\mathbb{Z}} \overline{f(x+\ell T')}\phi(x+\ell T' - kq)e^{ij\frac{2\pi}{T'}x}dx\right|^2 \\ &= T'^2\sum_{k\in\mathbb{Z}}\frac{1}{T'}\int_0^{T'}\left|\sum_{\ell\in\mathbb{Z}} \overline{f(x+\ell T)}\phi(x+\ell T - kq)\right|^2 dx.\end{aligned}$$

By the same proof as above, we also have (1.19). This completes the proof of the theorem.

2.3. Proof of Corollary 3. Let $\{\psi_{j,k}^*\}$ be the dual basis of L^2 relative to $\{\psi_{j,k}\}$. Then $\{\psi_{j,k}^*\}$ is also an unconditional basis of L^2, and it follows from (1.21) that

$$L^{-1}\sum_{j,k\in\mathbb{Z}} |a_{j,k}|^2 \leq \left\|\sum_{j,k\in\mathbb{Z}} a_{j,k}\psi_{j,k}^*\right\|^2 \leq K^{-1}\sum_{j,k\in\mathbb{Z}} |a_{j,k}|^2 \tag{2.12}$$

for all $\{a_{j,k}\} \in \ell^2(\mathbb{Z}^2)$. Now, for any $f \in L^2$, writing

$$f(x) = \sum_{j,k \in \mathbb{Z}} a_{j,k}\psi^*_{j,k}(x),$$

we have

$$a_{j,k} = \langle f, \psi_{j,k} \rangle,$$

so that (2.12) yields

$$L^{-1} \sum_{j,k \in \mathbb{Z}} |\langle f, \psi_{j,k} \rangle|^2 \le \|f\|^2 \le K^{-1} \sum_{j,k \in \mathbb{Z}} |\langle f, \psi_{j,k} \rangle|^2,$$

which is equivalent to

$$K\|f\|^2 \le \sum_{j,k \in \mathbb{Z}} |\langle f, \psi_{j,k} \rangle|^2 \le L\|f\|^2 \tag{2.13}$$

for all $f \in L^2$. Hence, (1.25) follows from Theorem 1 with $a = 2$ and $b = 1$. In addition, (1.26) follows from (1.11) for these specific values of a and b. Of course, (1.27) and (1.28) are now the analogous consequences of (2.12) with $\psi^*_{j,k}$ replaced by $\tilde{\psi}_{j,k}$, where $\tilde{\psi}$ is the dual of ψ.

2.4. Proof of Theorem 4. Before we go ahead to establish the theorem, let us first say something about the dual $\tilde{\psi}$ of ψ. In the first place, we note that $\{\psi_{0,k}\} = \{\psi(\cdot - k)\}$ is an unconditional basis of W_0, meaning, of course, that it is complete in W_0 and satisfies

$$K \sum_{k \in \mathbb{Z}} |b_k|^2 \le \left\| \sum_{k \in \mathbb{Z}} b_i \psi(\cdot - k) \right\|^2 \le L \sum_{k \in \mathbb{Z}} |b_k|^2 \tag{2.14}$$

for all $\{b_k\} \in \ell^2$, where $0 < K \le L < \infty$ are as in (1.21). Note that (2.14) is a consequence of (1.21) with the same bounds. It is well known (cf. [7]) that (2.14) is equivalent to

$$K \le \Psi(\omega) \le L \quad \text{a.e.} \tag{2.15}$$

for the same constants K and L, where $\Psi(\omega)$ was defined in (1.32). Hence, we have

$$\Psi(\omega) = \sum_{j \in \mathbb{Z}} c_j e^{-ij\omega}$$

for some $\{c_j\} \in \ell^2$, so that

$$\tilde{\psi}(x) = \sum_{j \in \mathbb{Z}} c_j \psi(x - j)$$

is in W_0. Secondly, from the definition of $\tilde{\psi}$ in (1.31), we see that

$$\sum_{k \in \mathbb{Z}} \hat{\tilde{\psi}}(\omega + 2\pi k)\overline{\hat{\psi}(\omega + 2\pi k)} = 1 \quad \text{a.e.}, \tag{2.16}$$

and by a standard argument, it can be shown that (2.16) is equivalent to

$$\langle \psi(\cdot - j), \tilde{\psi}(\cdot - k)\rangle = \delta_{j,k}, \qquad j,k \in \mathbb{Z}. \tag{2.17}$$

In the proof of Theorem 4, we orthonormalize $\{\psi_{j,k}\}$ as usual by defining a function η, with

$$\hat{\eta}(\omega) = \frac{\widehat{\psi}(\omega)}{\Psi(\omega)^{\frac{1}{2}}}. \tag{2.18}$$

In view of (2.15), we have $\eta \in W_0$, and

$$\eta_{j,k}(x) = 2^{\frac{j}{2}}\eta(2^j x - k), \qquad j,k \in \mathbb{Z},$$

constitute an orthonormal basis of L^2. Hence,

$$\sum_{j,k} |\langle f, \eta_{j,k}\rangle|^2 = \|f\|^2, \qquad f \in L^2,$$

and by Corollary 3 with $K = L = 1$, we have

$$\sum_{j\in\mathbb{Z}} |\hat{\eta}(2^j\omega)|^2 = 1 \quad \text{a.e.}, \tag{2.19}$$

so that (1.33) follows from (1.31), (1.32), (2.18), and (2.19). Furthermore, (1.34) is a consequence of (1.33) by following the same method of derivation of (1.11). This completes the proof of the theorem.

2.5. Proof of Theorem 5. By (1.39), we have

$$\overline{\widehat{\psi}(2\omega)}\ \widehat{\widetilde{\psi}}(2\omega) = \overline{m_0(\omega+\pi)}\widetilde{m}_0(\omega+\pi)\overline{\hat{\phi}(\omega)}\ \hat{\tilde{\phi}}(\omega),$$

and hence it follows from (1.37) and (1.38) that

$$\begin{aligned}\overline{\widehat{\psi}(2\omega)}\ \widehat{\widetilde{\psi}}(2\omega) &= (1 - \overline{m_0(\omega)}\widetilde{m}_0(\omega))\overline{\hat{\phi}(\omega)}\ \hat{\tilde{\phi}}(\omega) \\ &= \overline{\hat{\phi}(\omega)}\ \hat{\tilde{\phi}}(\omega) - \overline{\hat{\phi}(2\omega)}\ \hat{\tilde{\phi}}(2\omega).\end{aligned} \tag{2.20}$$

Under the assumption on the decay property of ϕ and $\tilde{\phi}$, we note that $\hat{\phi}$ and $\hat{\tilde{\phi}}$ are both continuous at zero and converge uniformly to zero as $|\omega| \to \infty$. Therefore, by telescoping, we have, from (2.20),

$$\sum_{j\in\mathbb{Z}} \overline{\widehat{\psi}(2^j\omega)}\ \widehat{\widetilde{\psi}}(2^j\omega) = \overline{\hat{\phi}(0)}\ \hat{\tilde{\phi}}(0) = 1$$

for all ω, where the assumption $m_0(0) = \widetilde{m}_0(0) = 1$ in (1.37) is used. This establishes (1.40). Since (1.41) follows from (1.40) as in the derivation of (1.11), we have completed the proof of the theorem.

3. Final remarks. In this paper we have established various inequalities and identities of Littlewood–Paley type, among which are those for the following three classes of functions:

(a) The collection $\mathcal{F}$ of all s.t. frames;

(b) The collection $\mathcal{R}$ of all ψ such that $\{\psi_{j,k}\}$, as defined in (1.20), is an unconditional basis of L^2;

(c) The subcollection $\mathcal{W}$ of functions in $\mathcal{R}$ that have duals as defined by (1.22)–(1.24).

It was shown in §2.3 that $\mathcal{R} \subset \mathcal{F}$, and, in fact, the bounds for the unconditional basis $\{\psi_{j,k}\}$ generated by any $\psi \in \mathcal{R}$ in (1.21) remain to be the frame bounds for the frame $\{\psi_{j,k}\}$. Hence, we have

$$\mathcal{W} \subset \mathcal{R} \subset \mathcal{F}. \tag{3.1}$$

Observe that even for any $\psi \in \mathcal{F}$, the result

$$\int_{-\infty}^{\infty} \frac{|\widehat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty$$

from (1.11) already yields

$$\int_{-\infty}^{\infty} \psi(x)dx = \widehat{\psi}(0) = 0,$$

provided that $\widehat{\psi}$ is continuous at zero, which follows from a very weak growth condition of ψ, such as $\psi \in L^1(-\infty,\infty)$. Hence, it is already reasonable to call any $\psi \in \mathcal{F}$ a "wavelet." If, in addition, $\psi \in \mathcal{W}$ with dual $\widetilde{\psi}$, then by using the notion of "integral wavelet transform" as defined in (1.5), namely,

$$(W_j f)(x) = 2^j \int_{-\infty}^{\infty} f(y)\overline{\psi(2^j y - 2^j x)}dy,$$

the coefficients

$$a_{j,k} := 2^{\frac{-j}{2}}(W_j f)\left(\frac{k}{2^j}\right), \tag{3.2}$$

of the "wavelet series" expansion

$$f(x) = \sum_{j,k\in\mathbb{Z}} a_{j,k}\widetilde{\psi}_{j,k}(x) \tag{3.3}$$

of any $f \in L^2$, contain very important information of the "signal" f in time-frequency analysis.

Regarding (3.1), while it is obvious that $\mathcal{R}$ is a proper subset of $\mathcal{F}$, it is not immediately clear that $\mathcal{W}$ is properly contained in $\mathcal{R}$. In the following, we give a very simple proof of this fact.

Let $\psi \in L^2$ be such that $\{\psi_{j,k}\}$, as defined in (1.20), is any orthonormal basis of L^2, and consider the function

$$\eta(x) := \psi(x) - \bar{z}2^{\frac{1}{2}}\psi(2x),$$

where z is any complex parameter with $|z| < 1$. Then it is clear that $\eta \in \mathcal{R}$. In fact, we have

$$(1-|z|)^2 \sum_{j,k\in\mathbb{Z}} |a_{j,k}|^2 \le \left\| \sum_{j,k\in\mathbb{Z}} a_{j,k}\eta_{j,k} \right\|^2 \le (1+|z|)^2 \sum_{j,k\in\mathbb{Z}} |a_{j,k}|^2$$

for any $\{a_{j,k}\} \in \ell^2(\mathbb{Z}^2)$. Let $\{\eta^*_{j,k}\}$ be the basis dual to $\{\eta_{j,k}\}$. Then it is easy to verify that

$$\text{(3.4)} \qquad \begin{aligned} \eta^*_{0,0}(x) &= \sum_{\ell=0}^{\infty} z^\ell \psi_{-\ell,0}(x), \\ \eta^*_{0,1}(x) &= \psi_{0,1}(x). \end{aligned}$$

Hence, for η to be in $\mathcal{W}$, the family $\{\eta^*_{j,k}\}$ must be given by

$$\eta^*_{j,k}(x) = 2^{\frac{j}{2}}\eta^*_{0,0}(2^jx-k),$$

and in particular,

$$\eta^*_{0,1}(x+1) = \eta^*_{0,0}(x),$$

so that (3.4) yields

$$\psi(x) = \eta^*_{0,1}(x+1) = \eta^*_{0,0}(x) = \psi(x) + \sum_{\ell=1}^{\infty} z^\ell \psi_{-\ell,0}(x),$$

or

$$\sum_{\ell=1}^{\infty} \psi_{-\ell,0}(x) z^\ell = 0.$$

This is not possible, unless $\psi_{-\ell,0}(x) = 0$ or $\psi(x) = 0$ for all x.

The consideration of $\eta^*_{0,0}$ in (3.4) was motivated by the fundamental work of Daubechies [5, p. 989], where a corresponding function $h^*_{0,0}$ was constructed by using the Meyer wavelet (cf. [10]). This function $h^*_{0,0}$ is not in $L^p(-\infty,\infty)$ for sufficiently small $p-1>0$ as shown by our derivation and (3.4) above. Note that this corrects a mistake in [5], where it was erroneously stated that $h^*_{0,0}$ was not in $L^p(-\infty,\infty)$ for large p. On the other hand, the fact that $\mathcal{W} \neq \mathcal{R}$ also follows from an earlier result of Tchamitchian [11], [12], as pointed out by Daubechies [5, p. 989] and Meyer [10, p. 127].

Acknowledgment. We are grateful to the referees for their valuable comments.

REFERENCES

[1] C. K. Chui and J. Z. Wang, *On compactly supported spline-wavelets and a duality principle*, Trans. Amer. Math. Soc., 330 (1992), pp. 903–915.

[2] ———, *A general framework of compactly supported splines and wavelets*, J. Approx. Theory, to appear.

[3] A. Cohen, *Ondelettes, Analyses multiresolutions, et traitement numerique du signal*, Ph.D. thesis, University Paris-Dauphine, 1990.

[4] A. Cohen, I. Daubechies, and J. C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., to appear.

[5] I. Daubechies, *The wavelet transform, time-frequency localization and signal analysis*, IEEE Trans. Inform. Theory, 36 (1990), pp. 961–1005.

[6] R. J. Duffin and A. C. Schaeffer, *A class of nonharmonic Fourier series*, Trans. Amer. Math. Sol., 72 (1952), pp. 341–366.

[7] S. G. Mallat, *Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$*, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.

[8] S. G. Mallat and W. L. Hwang, *Singularity detection and processing with wavelets*, preprint, 1991.

[9] S. G. Mallat and S. Zhong, *Wavelet transform maxima and multiscale edges*, in Wavelets and Their Applications, M. B. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, and L. Raphael, eds., Jones and Bartlett, Boston, 1992, pp. 67–104.

[10] Y. Meyer, *Ondelettes*, Vol. 1, Hermann, Paris, 1990.

[11] Ph. Tchamitchian, *Cacul symbolique sur les opérateurs de Caldéron-Zygmund et bases inconditionnelles de $L^2(\mathbb{R}^n)$*, C.R. Acad. Sci. Paris, 303 (1986), pp. 215-218.

[12] ———, *Biorthogonalité et théorie des opérateurs*, Rev. Mat. Iberoamericana, to appear.

SIAM J. MATH. ANAL.
Vol. 24, No. 2, pp. 279-289, March 1993

CONSERVATION LAWS WITH DISCONTINUOUS FLUX FUNCTIONS*

TORE GIMSE†

Abstract. The author studies the initial value problem for the scalar conservation law $u_t + f(u)_x = 0$ in one spatial dimension. The flow function may be discontinuous with a finite number of jump discontinuities. This paper proves existence of a weak solution, and the proof is constructive, suggesting a numerical method for the problem.

Key words. conservation laws, discontinuous flow, front tracking, porous media

AMS(MOS) subject classifications. 35L65, 35L67, 76T05

Introduction. In this paper we are interested in the Cauchy problem for the scalar conservation law

$$u_t + f(u)_x = 0, \tag{0.1}$$

that is, the initial value problem with $u(x, 0) = u_0(x)$ piecewise continuous of bounded variation, and so that $f_0(x) = f(u_0(x))$ has bounded variation.

The flux function f is supposed to be piecewise smooth with a finite number of jump discontinuities. For simplicity, we will consider flux functions with only one point of discontinuity, so that

$$\lim_{u \to \bar{u}^-} f(u) \neq \lim_{u \to \bar{u}^+} f(u),$$

$\bar{u}$ being the point of discontinuity. The extension to a finite number of discontinuities is outlined at the end of the paper.

This Cauchy problem may arise in several physical applications. For two phase flow in porous media we may have a discontinuous flux (flow) function if the flow properties change abruptly at some saturation. Such changes are obtained for the relative permeability at the irreducible saturation, both when measuring the relative permeability experimentally [11], [16], and when modeling flow properties on a network of pores [12]. This effect is due to discontinuous distribution of the low saturation, and is a jump from zero permeability value to a presumably small but positive value at this critical saturation. Simulations on discretized fracture apertures indicate possible major discontinuities for the nonwetting phase relative permeability, particularly for systems with small long-range correlation among apertures in the direction of the flow [19]. A discontinuity of the relative permeability yields a corresponding jump for the flow function. In standard texts of reservoir simulation and related topics, e.g., [4], relative permeability curves are assumed to be continuous, or approximated by continuous functions. This paper, however, suggests that also discontinuous functions, which in some cases may be more realistic, may be used with existence and stability results similar to those for the continuous problem.

It should be an object of further investigation if our results could be extended to be applicable also for hysteresis problems, that is, history dependent flow properties. Laboratory studies [6] indicate that we would expect to have an interval of saturations, say (u_1, u_2) where $f(u)$ is double valued, and the correct flow value is determined by

* Received by the editors February 12, 1992; accepted for publication July 14, 1992. This author's research was supported by the Royal Norwegian Council for Technical and Industrial Research.

† Department of Mathematics, University of Oslo, N-0316, Oslo, Norway.

previous or neighboring saturation values. Marchesin et al. [17] have studied this problem, but their analysis is based on finite slopes of flow functions.

Another possible application is traffic flow analysis [15]. We propose the following model for two-lane unidirectional traffic on a freeway that involves a discontinuous flow function: Assume that all cars have the same length, and that the speed of cars in the left lane is constant, independent of the car density (at least at those values of interest here). In the right lane, a certain fraction of the cars drive with a low, fixed speed, but passing (by changing lanes only during passing, and with instantaneous acceleration) is possible. Thus, as long as the density in the left lane permits passing, that is, as long as there is space enough between the cars, the overall flow depends continuously upon the overall density. However, as the density reaches the value where the left lane density prohibits passing, the overall flow drops discontinuously to that of the two lanes considered separately. Although multilane traffic with passing has been studied previously (e.g., [18]), no model similar to the one proposed above is known by the author. The consequences of this model should be an object of future investigation.

In either application, the procedures of this paper are constructive, and suggest a numerical method. The major idea of our method is to approximate the flux function f with a piecewise linear function, and approximate the initial value function u_0 with a step function [3], [7], [9]. By this procedure the original Cauchy problem is approximated by Riemann problems, and the solution of these consists of shocks only. We call this method a front tracking method. Shocks of the solution are traced without numerical dispersion, whereas rarefaction waves are approximated by a sequence of small shocks. Such methods have been extensively developed by the Oslo group [1], [2], and have turned out to be computationally and mathematically successful.

The following definition simplifies the notation.

DEFINITION. Let u_- and u_+ denote the points $(\bar{u}, \lim_{u\to\bar{u}^-} f(u))$ and $(\bar{u}, \lim_{u\to\bar{u}^+} f(u))$, respectively. We write $u_- \prec u_+$ if $\lim_{u\to\bar{u}^-} f(u) < \lim_{u\to\bar{u}^+} f(u)$, and say that f is double valued at the jump discontinuity at $u = \bar{u}$.

Throughout this paper we will assume that $u_- \prec u_+$.The case $u_+ \prec u_-$ can be treated symmetrically. We will treat u_- and u_+ as being two different u values, and we will let $\bar{u}$ denote any of them.

The fact that f is discontinuous implies that the existence results of, e.g., Krushkov [13] and Kuznetsov [14] do not apply to this problem. A somewhat similar "discontinuous problem" is the problem with a flux function discontinuously varying with x. This latter problem is solved in [5] by combining a technique of Temple [20] with front tracking methods by Dafermos [3] and Holden, Holden and Høegh-Krohn [9]. In this paper we will build mainly on [9]. By using front tracking as our method of analysis, we can avoid estimates involving the boundedness of the derivative of f, and thereby we are able to prove existence of a solution of the Cauchy problem. Our work will be based on, extend, and partly parallel the previous works by Holden, Holden, and Høegh-Krohn [8], [9], where similar techniques are used to study the continuous case. As for their works, our method is based on the solution of Riemann problems for (0.1), which will be discussed in some detail.

1. The solution of the Riemann problem. In general, the Riemann problem of (0.1) is the initial value problem consisting of two constant states separated by a discontinuity,

$$u(x,0) = \begin{cases} u_l & \text{for } x < 0, \\ u_r & \text{for } x > 0. \end{cases} \tag{1.1}$$

The Riemann problem when neither of u_l, u_r equals $\bar{u}$ is easily solved by the well-known procedure of taking convex envelopes of f between u_l and u_r. Note that even though f is not continuous between u_l and u_r, the convex envelope of f with respect to the interval (u_l, u_r) is continuous and piecewise smooth. Thus, we obtain the familiar fan-like solution picture in the $x-t$ plane of waves propagating with finite speed. In general the waves are smooth (rarefaction waves) or shocks, the latter being discontinuities traveling with a certain shock speed. A shock wave with left and right states u_1 and u_2 will be denoted a u_1/u_2 shock. However, since we may have, for example, $u_l < \bar{u} < u_r$, then, if $\bar{u}$ is part of the solution, we should specify whether we have u_- or u_+.

Special care should be taken when either u_l or u_r equals $\bar{u}$. The following lemma is easily verified by examining convex envelopes.

LEMMA 1.1. *The Riemann problem with initial values $u_l = \bar{u}$ and $u_r \neq \bar{u}$ has a unique solution with waves of finite speed only.*

However, if $u_r = u_- < u_l$, or $u_r = u_+ > u_l$, we have to extend the concept of convex envelopes.

DEFINITION. The convex envelope of the function f with respect to the interval (u_l, u_+), where f is double valued at $u_- < u_+$, is defined by the convex envelope of f with respect to the interval (u_l, u_-) connected to the line from u_- to u_+.

The convex envelope defined above is a curve in the $u-f(u)$ space, which may have infinite slope with respect to u. Thus, a general Riemann problem (1.1) is solved by tracing the convex envelopes of f with respect to the interval (u_l, u_r), using the definition above if necessary. The solution generally consists of a fan of waves with finite speed, and possibly one shock u_-/u_+ or u_+/u_- with infinite speed. Note that the Riemann problem $u_l = u_-$, $u_r = u_+$ or vice versa, is solved by a single shock of infinite speed to the right in the $x-t$ plane. However, since the u value is constant across such a shock, we call it a zero shock. Thus, in the sense of u, a zero shock carries no information, but the flux value information is transported instantaneously.

See Fig. 1.1 for a simple example of a Riemann problem solution.

2. Shock interactions. After the Riemann problem solution is found, we want to study the interaction of several Riemann problems. We will be particularly interested in the case of a piecewise linear flux function f, which implies that the only waves present are shocks [3], [9]. We define a single collision to be a collision involving and creating waves of finite speed only. That is, two or more waves interact at some point (x, t), none of which has infinite speed, and the result contains no zero shock.

In the following, rightnext front (or Riemann problem) will mean the next front (Riemann problem) to the right of the present. Starting out with finitely many Riemann

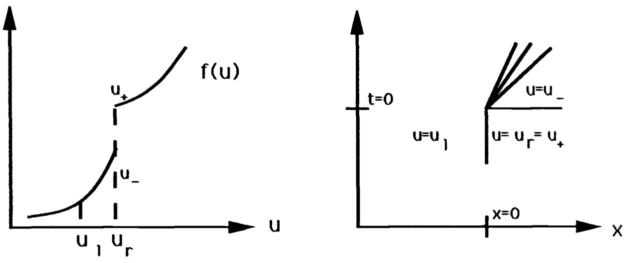


FIG. 1.1. *Discontinuous flux function and corresponding Riemann problem solution.*

problems as our initial data, we define the following algorithmic procedure for determining the solution $u(x, t)$:

(1) Solve the initial Riemann problems, starting from the left along the x axis. If a zero shock evolves, change the left state of the rightnext Riemann problem before solving that problem;

(2) After having finished at $t=0$, determine the first interaction to occur, say, at $t=\tau$. Denote the interacting constant states by $u_1, u_2, \dots, u_M$, $M>1$. Here u_1 is the leftmost state and u_M is the rightmost. The interaction is resolved by solving the Riemann problem with initial values $u_l=u_1$ and $u_r=u_M$. If a zero shock occurs, an interaction is created instantaneously at the rightnext front. If this happens, or if more interactions occur at the same time, treat them from the left, while changing the corresponding left value of the next front when zero shocks appear;

(3) When all interactions at τ are resolved, proceed to the next interaction at some greater time, etc.

As discussed above, a created zero shock, of course, will influence the rightnext front (or the rightnext interaction), but the following lemma assures limited distribution.

LEMMA 2.1. *A zero shock emerging from* (x, t) *interacts with the rightnext front instantaneously, but only with the rightnext.*

Proof. Assume that a u_-/u_+ shock is formed. The rightnext front is necessarily of the kind $u_+/\tilde{u}$ where $\tilde{u}\neq\bar{u}$, since if $\tilde{u}=u_+$ we had no front, and if $\tilde{u}=u_-$ the rightnext front was a zero shock as well, which is impossible since our resolution starts from the left. Thus, the rightnext front is turned into a Riemann problem with initial states u_- and $\tilde{u}$, which, by Lemma 1.1 is solved by shocks of finite speed only. A similar argument is valid if the zero shock is a u_+/u_- shock. □

If there is an interaction rightnext to the true collision, influence further to the right depends on the right state of that interaction.

Provided we have a finite number of interactions at time t, this completely resolves and continues the solution. It remains to be determined whether this procedure of resolution is well defined, that is, whether the solution is independent of the order in which simultaneous interactions are resolved. Firstly, by Lemmas 1.1 and 2.1, it is easily seen that the only cases that need to be checked are when more interactions occur with no fronts between them. The following lemma determines the resulting solution of a sequence of simultaneous interactions.

LEMMA 2.2. *For any finite sequence of simultaneous interactions creating zero shocks, the overall result is determined by the leftmost interaction.*

Proof. Let $I_1, I_2, \dots, I_N$ be the sequence of simultaneous interactions. Note that the left state of I_1 and the right state of I_N may be different from $\bar{u}$, but that the rest of the left and right states involved equal u_- and u_+ alternately. We will demonstrate that the order in which the interactions are treated does not affect the overall solution. Assume that the sequence of Is is already obtained, and that the next interaction to consider is I_1. Assume that the right state of I_1 is u_-. The case of $u_r=u_+$ is treated symmetrically. Thus, by the resolution of I_1, a u_+/u_- shock evolves, changing the left state of I_2 to u_+. However, by our assumption of the sequence, the right state of I_2 was u_+, so that the interaction I_2 is killed. The next interaction is not altered, and I_3 now is a new leftmost interaction in the remaining sequence. Thus, by continuing this argument, we see that the entire sequence is resolved by a u_+ state, which was determined by the u_+/u_- shock emerging from the leftmost interaction. □

Thus the resolution procedure defined by treating interactions with increasing x is well defined. We will define an event to be either a single collision or one or more simultaneous interactions, each creating zero shocks as described in Lemma 2.2. The latter will be denoted as a dual collision. See Fig. 2.1 for different kinds of events.

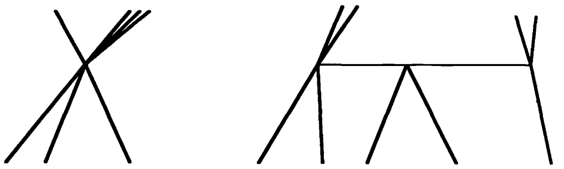


FIG. 2.1. *Single collision* (*left*) *and more interactions* (*right*).

Having determined the well-defined algorithm for treating Riemann problems locally, we are now able to examine the procedure of solving a finite number of initial Riemann problems globally as $t \to \infty$. The following theorem extends a result from [9].

THEOREM 2.3. *Given a piecewise linear flow function with one point of discontinuity, and an initial value function $u_0(x)$ consisting of finitely many constant states separated by discontinuities. Then, even for infinite time, only a finite number of events occur, and the overall solution $u(x, t)$ consists of a finite number of constant states, separated by shocks.*

Proof. Let N be the number of u values between which f is linear, plus the number of initial u values not in this set. Thus we may number the possible u values $w_1, w_2, \ldots, w_N$. Let $L(t)$ be the number of shock lines for $u(x, t)$, that is, the number of shock lines for a front w_i/w_j is $|i-j|$, and let $F(t)$ be the number of shocks in $u(x, t)$. Define the function $G(t) = NL(t) + F(t)$. Then $G(t)$ is obviously nonnegative. We will show that $G(t)$ is decreasing at each event, leaving us with a finite number of possible events only. First, if the event is a true collision, the theorem from [9] is valid. Examine, therefore, a dual collision. We will compare the dual collision with two collisions, connected by a zero shock of large but finite speed (see Fig. 2.2). Note that we may always find a speed S so that no other interaction takes place before the shock with speed S reaches the position of the right interaction. We name this the split case. Note that the result in the two cases are the same. Obviously, G_{before} and

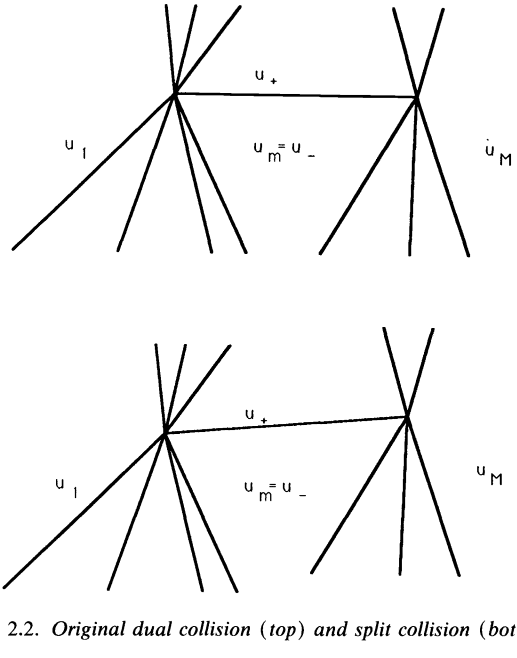


FIG. 2.2. *Original dual collision* (*top*) *and split collision* (*bottom*).

G_{after} is the same for the two cases, and since we know that G is decreasing for the split case [9], the same is valid for the dual collision. If more intermediate interactions were killed in between the left and right interaction, it is easily seen that G decreases even more. Thus, we have a finite number of interactions, which gives only a finite number of shocks, dividing the $x-t$ plane in a finite number of polygons where the solution u is constant. □

COROLLARY. *The total variation of the solution is nonincreasing.*

3. Stability. We now turn our interest to the stability of the solution, both with respect to $u_0(x)$ and the flux function $f(u)$. The following theorem ensures stability with respect to the initial data.

THEOREM 3.1. *If $u(x, t)$ and $v(x, t)$ solve* (0.1) *with initial value functions $u_0(x)$ and $v_0(x)$, respectively, u_0 and v_0 being step functions with finitely many values, and so that $u_0(x)=v_0(x)$ outside some finite interval $[-a, a]$, and f being piecewise linear with one point of discontinuity, then*

$$\int |u(x, t)-v(x, t)|\, dx \leqq \int |u_0(x)-v_0(x)|\, dx.$$

Proof. Assume that $u_0(x)$ and $v_0(x)$ are constant at the intervals $I_i=(a_i, a_{i+1})$, where $i=1,2,\ldots, M$, and $a_1=-\infty$, $a_{M+1}=\infty$. We want to construct a sequence $\{u_{0,n}\}_{n=1}^N$ so that $u_{0,1}=u_0$ and $u_{0,N}=v_0$. This construction is done by taking the intervals I_i one by one, and by moving the previous $u_{0,k}$ towards v_0 at $\frac{1}{3}$ of an interval every time. Thus, if $u_0=w_{s_i}$ and $v_0=w_{t_i}$ at interval I_i, then $N=\sum_{i=1}^M 3|s_i-t_i|$. Let $\{w_j\}$ be the set of initial and possible values for u. Note that $u_{0,i}$ differs from $u_{0,i+1}$ only at $\frac{1}{3}$ of some interval I_k, and that $|u_{0,i}-u_{0,i+1}|=|w_j-w_{j+1}|$ for some j at this interval. Furthermore, $|u_0(x)-v_0(x)|_{L_1}=\sum_{i=1}^{N-1}|u_{0,i}-u_{0,i+1}|_{L_1}$. Let $u_i(x, t)$ be the solution of (0.1) with initial value $u_{0,i}$. We then have

$$\begin{aligned}\int |u(x, t)-v(x, t)|\, dx &\leqq \int \sum_{i=1}^{N-1} |u_i-u_{i+1}|\, dx \\ &\leqq \int \sum_{i=1}^{N-1} |u_{0,i}-u_{0,i+1}|\, dx = \int |u_0(x)-v_0(x)|\, dx,\end{aligned}$$

the latter inequality by Lemma 3.2 below that is taken from [8]. □

LEMMA 3.2 (Holden, Holden, and Høegh-Krohn).

$$\int \sum_{i=1}^{N-1} |u_i-u_{i+1}|\, dx \leqq \int \sum_{i=1}^{N-1} |u_{0,i}-u_{0,i+1}|\, dx.$$

Proof. The proof [8] considers the time derivative of $\int |u_i-u_{i+1}|\, dx$ at the intervals from Theorem 3.1. To transfer the result from [8], we observe that this derivative is zero also if $u_i=u_-$ and $u_{i+1}=u_+$ or vice versa. □

Note that Theorem 3.1 implies stability also for higher-dimensional problems. This follows by the dimensional splitting analysis by Holden and Risebro [10].

Next we are interested in stability with respect to the flux function f. At this point we will assume that the discontinuity of f is fixed, and so are the two corresponding points u_- and u_+. With this assumption, we may state the theorem.

THEOREM 3.3. *Let f and g be piecewise linear functions with a coinciding point of discontinuity at $u=\bar{u}$, and let $v(x, t)$ and $u(x, t)$ be the corresponding solutions of $u_t+f(u)_x=0$ and $v_t+g(v)_x=0$ with the same initial value, a step function taking finitely* *many values*: $u_0(x)=v_0(x)$. *Then*

$$\frac{d}{dt}\int |u(x,t)-v(x,t)|\,dx \leqq TV_x(f(u_c(x,t))-g(v_c(x,t)))$$
$$\leqq TV_x(f(u_{0,c}(x,t))-g(v_{0,c}(x,t))),$$

where $u_c(x,t)$ *and the Total Variation* (TV_x) *are defined below.*

DEFINITION. Let u_i be the value of the step function $u(x,t)$ taken at the interval (a_i, a_{i+1}), $i=1,2,\dots,M$, for fixed t. Then $u_c(x,t)$ is defined by

$$u_c(x,t)=\begin{cases} u_i & \text{for } a_i \leqq x \leqq a_{i+1}-\varepsilon, \\ u_i+\dfrac{(x-a_{i+1}+\varepsilon)}{\varepsilon}(u_{i+1}-u_i) & \text{for } a_{i+1}-\varepsilon \leqq x \leqq a_{i+1}. \end{cases}$$

Here $\varepsilon=\frac{1}{3}\min_i\{a_{i+1}-a_i\}$.

Note that $u_c(x,t)$ is a piecewise linear, continuous function.

DEFINITION. $TV_x(f(u(x)))$ is defined by

$$TV_x(f(u(x)))=\sup\sum_{i=1}^{N}|f(u(x_{i+1}))-f(u(x_i))|,$$

where the supremum is taken over all finite partitions of $\{x_i\}$.

Note that u in the above definition should be continuous.

Proof of Theorem 3.3. The proof of Theorem 3.3 carries over literally from [8] by the following observation. Define the function $F(u)=f(u)-g(u)$, and note that since f and g are assumed to have identical discontinuities, F is continuous and piecewise linear. The analysis of [8] is based on estimates of $f-g$, and these estimates are still valid by the properties of F. □

We now have stability results for piecewise linear flux functions with piecewise constant initial data, and we will use this, together with knowledge of zero shocks, to conclude with existence and uniqueness results for problem (0.1).

4. Existence and uniqueness. We first restate the problem that will be our object of study for the rest of this paper. The equation is

$$u_t+f(u)_x=0, \tag{4.1}$$

with initial data $u(x,0)=u_0(x)$. The flux function f is measurable and continuous with bounded derivative, except at $u=\bar{u}$ as above. The initial value function $u_0(x)$ is measurable and of bounded variation, as is $f_0(x)=f(u_0(x))$. We assume there are values $u_s<\bar{u}<u_S$ and $x_s<x_S$, so that for $x\leqq x_s$, and $x\geqq x_S$, $u_0(x)$ is not in the interval (u_s, u_S). The latter restriction is put on $u_0(x)$ to avoid zero shocks traveling unlimited distances instantaneously. We have the following lemma to ensure this.

LEMMA 4.1. *There exist numbers s and S, $-\infty<s<S<\infty$, and so that for $x<x_s+st$ and $x>x_S+St$ we have either $u(x,t)<u_s$, or $u(x,t)>u_S$. In these areas the solution $u(x,t)$ is determined by the existence and uniqueness results in* [8].

Proof. Since $u_0(x)$ is of bounded variation, we may assume that x_S is so that either $u_0(x)<u_s$ or $u_0(x)>u_S$ for $x>x_S$, and similarly for $x<x_s$. The maximum speed of waves entering the region $x>x_S$ is then determined by the maximum slope of the function

$$f_S(u)=\begin{cases} f(u) & \text{for } u\leqq u_-, \\ f(u_-)+\dfrac{f(u_S)-f(u_-)}{u_S-u_-}(u-u_-) & \text{for } u_-\leqq u\leqq u_S, \\ f(u) & \text{for } u\geqq u_S. \end{cases}$$

By definition f_S has a finite maximum slope, S. Similarly we define f_s for waves entering the other region, $x < x_s$, and the lemma follows. □

Before proceeding we need the following lemma from [8].

LEMMA 4.2. *Assume that a measurable function f is approximated by a sequence of measurable, uniformly bounded functions $\{g_n\}$ satisfying*

$$|g_n(x) - f(x)| < \frac{1}{na_n} \quad \textit{for } x \in (a, b) - A_n,$$

where the Lebesgue measure of A_n, $m(A_n)$ satisfies $m(A_n) < 1/na_n$, $\{a_n\}$ being an increasing sequence of real numbers. Then for $m > n$, the sequence $\{g_n\}$ satisfies the following Cauchy criterion:

$$\int_a^b |g_n(x) - g_m(x)|\, dx \leqq \frac{2(b-a)}{na_n} + \frac{4M}{na_n},$$

where M is such that $|g_n(x)| < M$.

We are now in the position of constructing a sequence of solutions, which we will show converges to a solution of (4.1): For given k, we select k different u values, say $w_1, w_2, \ldots, w_k$, among which we should have the two entries for $\bar{u}$, u_-, and u_+. Then, for given f, we construct f_k by evaluating f at the chosen u values, making f_k piecewise linear between these values. Note that we by this construction keep the correct discontinuity. Finally we make a piecewise constant approximation of $u_0(x)$ from below, using only the k different u values at a finite number of sample points. We denote this approximation $u_{0,k}(x)$. Now, let $u_k(x, t)$ be the solution of the equation $u_t + f_k(u)_x = 0$ with initial data $u_{0,k}(x)$. This defines a sequence of solutions, and we have the following lemma.

LEMMA 4.3. *$\{u_i(x, t)\}$ is a Cauchy sequence in $L_{1,\mathrm{loc}}$.*

Proof. By the definitions made above, we apply Theorem 3.1,

$$\begin{aligned}&\int |u_i(x, t) - u_j(x, t)|\, dx \\ &\qquad \leqq \int |u_{0,i}(x) - u_{0,j}(x)|\, dx + tTV_x(f_i(u_{0,i,c}(x)) - f_j(u_{0,j,c}(x))).\end{aligned}$$

As for the corresponding result in [8] the right-hand terms vanish; the first by Lemma 4.2, and the second by the construction of f_i. Note that all f_i have the same discontinuity at $\bar{u}$, and are continuous elsewhere. Thus, the function F_{ij} defined by $F_{ij}(u) = f_i(u) - f_j(u)$ is continuous, which makes the second term vanish [8]. □

Since f is double valued at $u = \bar{u}$, we cannot conclude from Lemma 4.3 that the sequence of fluxes, $\{f_i(u_i)\}$ converges. However, by the knowledge of the Riemann problem solution we find the following lemma.

LEMMA 4.4. *If the original $u_0(x)$ is continuously increasing at x_0, where $u_0(x_0) = \bar{u}$, then for large i the approximated solution contains u_-, and vice versa.*

Proof. Since u_0 is continuously increasing, for i sufficiently large, the approximation $u_{0,i}$ is also increasing at x_0. Thus, the Riemann problem solution of convex envelopes invokes u_- but not u_+. □

LEMMA 4.5. *$\{f_i(u_i)\}$ is a Cauchy sequence in $L_{1,\mathrm{loc}}$.*

Proof. By Lemma 4.3 we know that $\{f_i(u_i)\}$ is Cauchy with respect to domains where $\{u_i\}$ is not converging to $\bar{u}$. Thus, it is sufficient to examine initial values close to $\bar{u}$. This is a study of cases, of which the continuously monotone cases are covered by Lemma 4.4. The remaining are true Riemann problems, of which we may have only finitely many (by the restrictions of u_0 and f_0), and by the Riemann problem solution algorithm, we have convergence also for these. □

We may now define the limiting functions of $\{u_i\}$ and $\{f_i(u_i)\}$ by defining the limit $u(x, t)$ to be the limit of $u_i(x, t)$ so that $f_i(u_i(x, t)) \to f(u(x, t))$. Note that this is a valid definition since by Lemma 4.3 we may define a family $\{\tilde{u}(x, t)\}$ so that for all $\tilde{u}$ in this family, $u_i \to \tilde{u}$ in $L_{1,\text{loc}}$. The $\tilde{u}$s differ only at sets of zero measure, or with respect to u_-/u_+. Thus, as f is single valued, $f_i(u_i) \to f(u)$ in $L_{1,\text{loc}}$, and the problem where f is double valued is resolved by Lemma 4.5, and thereby defining which $\bar{u}$ value to give the flux value $f(\bar{u})$.

THEOREM 4.6. *The limiting solution* $u(x, t)$ *defined above is a weak solution of* (4.1), *that is,*

$$\int_0^T \int (u(x, t)\phi_t(x, t) + f(u(x, t))\phi_x(x, t))\, dx\, dt + \int u_0(x)\phi(x, 0)\, dx = 0$$

for all $\phi \in C_0^1$.

Proof. Since every $u_i(x, t)$ is a weak solution of $u_t + f_i(u)_x = 0$, we have

$$\left| \int_0^T \int (u(x, t)\phi_t(x, t) + f(u(x, t))\phi_x(x, t))\, dx\, dt + \int u_0(x)\phi(x, 0)\, dx \right|$$

$$= \left| \int_0^T \int ([u(x, t) - u_i(x, t)]\phi_t(x, t) + [f(u(x, t)) - f_i(u_i(x, t))]\phi_x(x, t))\, dx\, dt \right.$$

$$\left. + \int (u_0(x) - u_{0,i}(x))\phi(x, 0)\, dx \right|$$

$$\leqq \int_0^T \int (|u(x, t) - u_i(x, t)|\,|\phi_t(x, t)|$$

$$+ |f(u(x, t)) - f_i(u_i(x, t))|\,|\phi_x(x, t)|)\, dx\, dt$$

$$+ \int |u_0(x) - u_{0,i}(x)|\,|\phi(x, 0)|\, dx.$$

Now let $K = \max\{|\phi|, |\phi_t|, |\phi_x|\}$, and investigate each term of the above expression:

$$\int_0^T \int |u(x, t) - u_i(x, t)|\,|\phi_t(x, t)|\, dx\, dt \leqq K \int_0^T \int |u(x, t) - u_i(x, t)|\, dx\, dt \to 0,$$

and

$$\int |u_0(x) - u_{0,i}(x)|\,|\phi(x, 0)|\, dx \leqq K \int |u_0(x) - u_{0,i}(x)|\, dx\, dt \to 0,$$

by the definition of $u(x, t)$ and $u_{0,i}(x)$. Finally, by Lemma 4.5 and the definition of $u(x, t)$,

$$\int_0^T \int |f(u(x, t)) - f_i(u_i(x, t))|\,|\phi_x(x, t)|\, dx\, dt$$

$$\leqq K \int_0^T \int |f(u(x, t)) - f_i(u_i(x, t))|\, dx\, dt \to 0. \qquad \square$$

Having proved existence of a weak solution, it remains to prove uniqueness of the solution. By uniqueness we mean that the constructive approach using front tracking gives a unique limit solution.

THEOREM 4.7. *The weak solution defined from Theorem* 4.6 *is the unique limit of the constructed sequence of piecewise constant solutions with respect to* $L_{1,\text{loc}}$.

Proof. Assume that both $v(x, t)$ and $u(x, t)$ are weak solutions of (4.1) constructed by the front tracking method. Then

$$\begin{aligned}\Delta &= \int |u(x, t) - v(x, t)|\, dx \\ &\leqq \int |u(x, t) - u_i(x, t)|\, dx + \int |u_i(x, t) - v(x, t)|\, dx \\ &\leqq \int |u(x, t) - u_i(x, t)|\, dx + \int |u_{0,i}(x) - u_0(x)|\, dx + t \sum_I TV_x(f_i(u_{0,i,c}) - f(u_0)),\end{aligned}$$

the latter by Theorem 3.3 and $v_0(x) = u_0(x)$. The sum runs over intervals I where u_0 is continuous. Thus, by the definitions of $u_{0,i}(x)$, $u_{0,i,c}(x)$, $u_i(x, t)$, $v(x, t)$, f_i, and f, Δ vanishes as $i \to \infty$. □

5. Finitely many discontinuities. The extension to a flow function with finitely many discontinuities where the one-sided limits exist is straightforward by the observation that the zero shocks that may occur at each Riemann problem solution are well defined. By well defined, we mean that given u_l and u_r we may have only one zero shock traveling to the left, and one traveling to the right. By symmetry arguments, the results of this paper are valid for zero shocks traveling in both positive and negative directions. Zero shocks colliding at a dual collision are identical, and, therefore, the algorithmic procedure for solving multiple Riemann problems is still valid when being careful about changing the correct left and right states at neighboring fronts and interactions.

REFERENCES

[1] F. BRATVEDT, K. BRATVEDT, C. F. BUCHHOLZ, T. GIMSE, H. HOLDEN, AND N. H. RISEBRO, *Front tracking for petroleum reservoirs*, in Ideas and Methods in Mathematical Analysis, Stochastics, and Applications, Cambridge Univ. Press, London, 1992.

[2] F. BRATVEDT, K. BRATVEDT, C. F. BUCHHOLZ, H. HOLDEN, L. HOLDEN, AND N. H. RISEBRO, *A new front tracking method for reservoir simulation*, SPE Res. Eng., (Feb. 1992) pp. 107-116.

[3] C. M. DAFERMOS, *Polygonal approximations of solutions of the initial value problem for a conservation law*, J. Math. Anal. Appl., 38 (1972), pp. 33-41.

[4] R. E. EWING, ED., *The Mathematics of Reservoir Simulation*, Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1983.

[5] T. GIMSE AND N. H. RISEBRO, *Solution of the Cauchy problem for a conservation law with discontinuous flux function*, SIAM J. Math. Anal., 23 (1992), pp. 635-648.

[6] R. E. GLADFELTER AND S. P. GUPTA, *Effect of fractional flow hysteresis on recovery of tertiary oil*; SPE J. (Dec. 1980), pp. 508-520.

[7] G. W. HEDSTROM, *Some Numerical Experiments with Dafermos Method for Nonlinear Hyperbolic Equations*, Lecture Notes in Math., 267, Springer-Verlag, Berlin, New York, 1972, pp. 117-138.

[8] H. HOLDEN AND L. HOLDEN, *On scalar conservation laws in one dimension*, in Ideas and Methods in Mathematical Analysis, Stochastics, and Applications, Cambridge Univ. Press, London, 1992.

[9] H. HOLDEN, L. HOLDEN, R. HØEGH-KROHN, *A numerical method for first-order nonlinear scalar conservation laws in one dimension*, Comput. Math. Appl., 15 (1988), pp. 595-602.

[10] H. HOLDEN AND N. H. RISEBRO, *A fractional steps method for scalar conservation laws without the* CFL *condition*, Math. Comp. (Jan. 1993).

[11] M. HONARPOUR AND S. M. MAHMOOD, *Relative-permeability measurements: an overview*, J. Pet. Tech., 40 (1988), pp. 963-966.

[12] A. KANTZAS AND I. CHATZIS, *Network simulation of relative permeability curves using a bond correlated-site percolation model of pore structure*, Chem. Eng. Comm., 69 (1988), pp. 191-214.

[13] S. N. KRUSHKOV, *First order quasilinear equations in several independent variables*, Math. USSR-Sb., 10 (1970), pp. 217-243.

[14] N. KUZNETSOV, *Weak solutions of the Cauchy problem for a multi-dimensional quasilinear equation*, Mat. Zametki, 2 (1967), pp. 401-410.

[15] M. J. LIGHTHILL AND G. B. WHITHAM, *On kinematic waves. I. Flood movement in long rivers*, Proc. Roy. Soc., 229A (1955), pp. 281-316; II. *Theory of traffic flow on long crowded roads*, Proc. Roy. Soc., 229A (1955), pp. 317-345.

[16] B. B. MAINI AND T. OKAZAWA, *Effects of temperature on heavy oil-water relative permeability of sand*, J. Can. Pet. Tech., 26 (1987), pp. 33-41.

[17] D. MARCHESIN, H. B. MEDEIROS, AND P. J. PAES-LEME, *A model for two phase flow with hysteresis*, Contemp. Math., 60 (1987), pp. 89-107.

[18] P. G. MICHALOPOULOS, D. E. BESKOS, AND Y. YAMAUCHI, *Multilane traffic flow dynamics: some macroscopic considerations*, Transportation Res., Part B, 18B (1984), pp. 377-395.

[19] K. PRUESS AND Y. W. TSANG, *On two-phase relative permeability and capillary pressure of rough-walled rock fractures*, Water Resour. Res., 26 (1990), pp. 1915-1926.

[20] B. TEMPLE, *Global solution of the Cauchy problem for a class of* 2×2 *non-strictly hyperbolic conservation laws*, Adv. Appl. Math., 3 (1982), pp. 335-375.

SIAM J. MATH. ANAL.
Vol. 24, No. 2, pp. 290-298, March 1993

002

ASYMPTOTIC BEHAVIOR OF ONE-STEP COMBUSTION MODELS WITH MULTIPLE REACTANTS ON BOUNDED DOMAINS*

JOEL D. AVRIN†

Abstract. The author considers reaction-diffusion systems on bounded domains modeling one-step reactions with Arrhenius kinetics in cases where the fuel consists of several species. The author assumes zero Neumann boundary conditions for the mass fractions and it is shown that one mass fraction decays to zero while, generally, residual amounts of the other species remain. These amounts are calculated explicitly from spatial averages of the initial conditions and it is shown that only if certain precise conditions are met will all mass fractions decay to zero. If additionally the temperature satisfies zero Neumann boundary conditions or fixed positive Dirichlet boundary conditions, then the temperature's asymptotic behavior is explicitly calculated as well.

Key words. reaction-diffusion equations, boundary conditions, spatial averages, steady-state convergence

AMS(MOS) subject classifications. 35B40, 35K57, 80A25

1. Introduction. We consider one-step reactions with multiple-species fuels that arise in combustion theory. An example of this type of reaction is (see, e.g., [6, p. 6])

$$2NO+Cl_2 \rightarrow 2NOCl. \tag{1.1}$$

The general one-step reaction involving the reactants and products $A_1, \dots, A_N$ can be written as

$$\sum_{i=1}^{N} \nu_i A_i \rightarrow \sum_{i=1}^{N} \mu_i A_i \tag{1.2}$$

where the stoichiometric coefficients ν_i and μ_i are positive integers; $\mu_i = 0$ (respectively, $\nu_i = 0$) for each A_i that is not a product (respectively, not a reactant) in (1.2). In (1.1), for example, if $A_1 = NO$, $A_2 = Cl_2$, and $A_3 = NOCl$, then $\nu_1 = \mu_3 = 2$, $\nu_2 = 1$, while $\nu_3 = \mu_1 = \mu_2 = 0$.

By relabeling, if necessary, we can assume that for some M with $1 < M < N$ $A_1, \dots, A_M$ are the reactants, i.e., $\nu_i \neq 0$, $i = 1, \dots, M$. Let $Y_i = Y_i(x, t)$, $i = 1, \dots, M$ denote the respective mass fractions of the A_i and let $T = T(x, t)$ denote the (dimensionless) temperature. Here $x \in \Omega$, a bounded domain with smooth boundary $\partial\Omega$, and $t \geqq 0$. For positive constants B and E let $f(T)$ denote the Arrhenius rate law $f(T) = B \exp(-E/T)$, and set

$$\omega = \left[\prod_{i=1}^{M} Y_i^{\nu_i}\right] f(T). \tag{1.3}$$

For $i = 1, \dots, M$ set $\alpha_i = m_i \nu_i$, where m_i is the molecular mass of the ith species, then the following system of reaction-diffusion equations models the reaction (1.2) within the framework of the isobaric approximation of slow combustion:

$$T_t = \Delta T + Q\omega, \tag{1.4a}$$

$$Y_{it} = d_i \Delta Y_i - \alpha_i \omega, \qquad i = 1, \dots, M. \tag{1.4b}$$

* Received by the editors April 1, 1992; accepted for publication (in revised form) August 5, 1992.

† Department of Mathematics, University of North Carolina at Charlotte, Charlotte, North Carolina 28223.

Here Q, $d_1, d_2, \ldots, d_M$ are positive constants, with Q denoting the heat release. By way of example, the reaction-diffusion equations for the reaction (1.1) are, with $\omega = Y_1^2 Y_2 f(T)$,

$$T_t = \Delta T + Q\omega, \tag{1.5a}$$

$$Y_{1t} = d_1 \Delta Y_1 - 2m_1\omega, \tag{1.5b}$$

$$Y_{2t} = d_2 \Delta Y_2 - m_2\omega. \tag{1.5c}$$

For further physical background on (1.3) and (1.4), see, e.g., [6], [27], or [13, pp. 57-59].

There is a wide body of literature associated with the simpler one-step reaction $A \rightarrow B$ and its associated set of reaction-diffusion equations. With $\Omega = \mathbb{R}$, the existence of traveling-wave solutions has been considered in [5], [14], [15], [23], [25], [26] and their stability vs. instability analyzed in [7], [21], [24]. The general Cauchy problem with $\Omega = \mathbb{R}$ has been considered in [1], [13], [19] and a similar model describing platelet aggregation has been studied in [17], [18]. The general Cauchy problem when Ω is a bounded domain has been studied, for example, in [2], [9], [10], [16], [20]; see also the references contained therein. Our purpose here will be to extend the qualitative theory developed in [2] to the system (1.4). Our basic tools will consist of standard comparison principles, a lemma appearing in a paper by Fitzgibbon and Martin [8, Lemma 2.5], and some integral estimates which we develop here that generalize some arguments used in [3].

We consider the following sets of boundary conditions on T and the Y_i, $i = 1, \ldots, M$: either

$$\frac{\partial T}{\partial \nu} = \frac{\partial Y_i}{\partial \nu} = 0 \quad \text{on } \partial\Omega, i = 1, \ldots, M, \tag{1.6}$$

where ν is the outward normal on Ω, or for a sufficiently smooth function g on $\partial\Omega$,

$$T = g \quad \text{on } \partial\Omega, \tag{1.7a}$$

$$\frac{\partial Y_i}{\partial \nu} = 0 \quad \text{on } \partial\Omega, i = 1, \ldots, M. \tag{1.7b}$$

Under these conditions it is possible to establish the following global existence and uniqueness result; the proof, by appealing to the techniques developed in [22, Chap. 14], is only a slight generalization of [2, Thm. 2.4] and can therefore be safely omitted.

THEOREM 1.1. *Let* $T_0 \equiv T(x, 0)$ *and* $Y_{i0} \equiv Y_i(x, 0)$, $i = 1, \ldots, M$, *be nonnegative and in* $C(\bar{\Omega})$ *and let w be the solution of the Dirichlet problem*

$$\Delta w = 0 \quad \text{in } \Omega, \tag{1.8a}$$

$$w = g \quad \text{on } \partial\Omega. \tag{1.8b}$$

For a given integer $k \geqq 2$ *choose g smooth enough so that* $w \in C^k(\bar{\Omega})$. *Then for either set of boundary conditions* (1.6) *or* (1.7) *there exist unique nonnegative global strong solutions* T *and* Y_i, $i = 1, \ldots, M$, *of* (1.4) *such that* $T, Y_i \in C^j((0, +\infty); C^k(\Omega)) \cap C([0, +\infty); C(\bar{\Omega}))$ *for each* $j \geqq 1$.

Before stating our main result we first note a fairly standard fact about the Laplace operator on Ω. Let $W_i(t) = \exp(d_i \Delta)$ where Δ is equipped with zero Neumann boundary conditions. Then for any $h \in C(\bar{\Omega})$ we have that

$$\lim_{t \to \infty} W_i(t)h = h_{AV} \tag{1.9}$$

uniformly in x, where

$$h_{AV}=\frac{1}{|\Omega|}\int_{\Omega} h(x)\,dx. \tag{1.10}$$

See, e.g., [2, Prop. 3.1] for a proof of this very intuitive result.

If we assume the conditions (1.6) and set $G(x,t)=T(x,t)+(Q/\alpha_1)Y_1(x,t)$, set $G_0=G(x,0)$ and let $(G_0)_{AV}$ equal the right-hand side of (1.10) with h replaced by G_0, then by taking appropriate linear combinations in (1.4), integrating over Ω, and dividing by $|\Omega|$ we obtain

$$\frac{1}{|\Omega|}\int_{\Omega} G(x,t)\,dx=(G_0)_{AV}. \tag{1.11}$$

Equation (1.11) will be useful to us in what follows; additionally, (1.11) serves as a basis for establishing uniform bounds on T. Bounds on all the Y_i follow immediately, since it is easy to see from the maximum principle that $\|Y_i(\cdot,t)\|_\infty \leqq \|Y_{i0}\|_\infty$, $i=1,\ldots,M$, for all $t\geqq 0$. That there exists a constant M_T such that $\|(T(\cdot,t)\|_\infty\leqq M_T$ for all $t\geqq 0$ follows from (1.11), using previous work by a number of authors on a particular class of reaction-diffusion systems of which (1.4) is a special case; see, e.g., [11] and the references contained therein. If conditions (1.7) are assumed, a bound on T can be obtained by modification of these arguments, but in fact boundedness of T in this case will follow independently from our convergence results below.

Let $(Y_{i0})_{AV}$ equal the right-hand side of (1.10) with h replaced by Y_{i0}, $i=1,\ldots,M$. By relabeling, if necessary, we can assume that $(1/\alpha_1)(Y_{10})_{AV}=\min_{1\leqq i\leqq M}\{(1/\alpha_i)(Y_{i0})_{AV}\}$. Set $(Z_{i0})_{AV}=(1/\alpha_i)(Y_{i0})_{AV}-(1/\alpha_1)(Y_{10})_{AV}$, $i=2,\ldots,M$, then each $(Z_{i0})_{AV}$ is nonnegative. With these preliminaries in mind we are ready to state our main result.

THEOREM 1.2. *Let T_0, Y_{i0}, $i=1,\ldots,M$, and $(Z_{i0})_{AV}$, $i=2,\ldots,M$, be as above and suppose without loss of generality that all of the $(Z_{i0})_{AV}$ are nonnegative. Suppose under either of the conditions* (1.6) *and* (1.7) *that T_0 is not identically zero and if conditions* (1.7) *are assumed suppose that $g(x)>0$ for all $x\in\partial\Omega$. Then $Y_1(x,t)$ converges uniformly in x to zero as $t\to\infty$, while*

$$\lim_{t\to\infty} Y_i(x,t)=\alpha_i(Z_{i0})_{AV},\qquad i=2,\ldots,M \tag{1.12}$$

uniformly in x. If conditions (1.6) *are assumed, $T(x,t)$ converges uniformly in x to $(G_0)_{AV}$ as $t\to\infty$, while if conditions* (1.7) *are assumed, $T(x,t)$ converges uniformly in x to w as $t\to\infty$, where w is as in* (1.8).

Remarks. (a) Note that Theorem 1.2 provides a complete description of the asymptotic behavior of the system (1.4), subject to the boundary conditions (1.6) or (1.7), whenever g is everywhere positive. The physical interpretation of (1.12) is that a residual amount of A_i always remains unless $(Z_{i0})_{AV}$ is identically zero, $i=2,\ldots,M$. There is certainly a precedent for the positivity assumption on g: see, e.g., [9], [20], and [10, §§ 5.1, 6.1, and 10.2] wherein the condition $g\equiv 1$ is assumed in a model of the one-species fuel case.

(b) The convergence of the Y_i to their respective steady-states is eventually exponential in t if and only if $\nu_1=1$. If $\nu_1>1$, the eventual convergence is no faster than a negative power of t, as we will see in the proof below.

The spatial averages of T and the Y_i for $t\geqq 0$ will also play a key role in the proof of Theorem 1.2. Following [8, § 2] we introduce the following notation for these quantities:

$$\bar{T}(t)=\frac{1}{|\Omega|}\int_\Omega T(x,t)\,dx, \tag{1.13a}$$

$$\bar{Y}_i(t)=\frac{1}{|\Omega|}\int_\Omega Y_i(x,t)\,dx, \qquad i=1,\dots,M. \tag{1.13b}$$

The following lemma follows from the proof of [8, Lemma 2.5] with only slight modifications made to the arguments so that they apply to (1.4) as well as the systems considered in [8].

LEMMA 1.1. *Let $\bar{T}$ and $\bar{Y}_i$, $i=1,\dots,M$, be as in* (1.13). *Then under the conditions of Theorem* 1.1 *we have that*

$$\lim_{t\to\infty}\|Y_i(\cdot,t)-\bar{Y}_i(t)\|_\infty=0, \qquad i=1,\dots,M \tag{1.14}$$

and if conditions (1.6) *are assumed we also have that*

$$\lim_{t\to\infty}\|T(\cdot,t)-\bar{T}(t)\|_\infty=0. \tag{1.15}$$

We will, for completeness, give a proof of Lemma 1.1 in § 3 below. The proof of Theorem 1.2 will appear in § 2. We close this section by demonstrating that under the conditions of Theorem 1.2 there exist constants $\alpha>0$ and $t_1>0$ such that $T(x,t)\geqq\alpha$ for all x in Ω and all $t\geqq t_1$.

PROPOSITION 1.2. *Under the conditions of Theorem* 1.1 *and the boundary conditions* (1.6) *there exists α and t_1 as noted above whenever T_0 is not identically zero.*

Proof. We have by Theorem 1.1 that $T(x,t)$ and $Y_i(x,t)$, $i\in\{1,2,3\}$, are nonnegative for all $t\geqq0$. Hence $\omega(x,s)$ is nonnegative for all $x\in\Omega$ and all $s\geqq0$ by (1.3). Let $W_0(t)=\exp(td_0\Delta)$ where Δ is equipped with zero Neumann boundary conditions, then T satisfies the integral equation

$$T(t)=W_0(t)T_0+Q\int_0^t W_0(t-s)\omega(s)\,ds, \tag{1.16}$$

where we have suppressed the dependence on x. Since $W_0(t)$ preserves nonnegativity, it follows from (1.16) (or the maximum principle) that $T(x,t)\geqq(W_0(t)T_0)(x)$ for all $t\geq0$. The result then follows from (1.9) with h replaced by T_0, since if T_0 is not identically zero then $(T_0)_{AV}$ is strictly positive; here we can have $\alpha=(T_0)_{AV}-\varepsilon$ for any ε with $0<\varepsilon\ll(T_0)_{AV}$.

PROPOSITION 1.3. *Under the conditions of Theorem* 1.1 *and the boundary conditions* (1.7) *there exist α and t_1 as noted above for any g with $g(x)>0$ for all $x\in\partial\Omega$.*

Proof. This result is almost identical to [2, Thm. 4.2]; we include a proof for completeness. Again we have that $\omega(x,s)$ is nonnegative for all $x\in\Omega$ and all $s\geqq0$. The integral equation for T is now (for w as in (1.8))

$$T(t)=W_0(t)(T_0-w)+w+Q\int_0^t W_0(t-s)\omega(s)\,ds, \tag{1.17}$$

where now Δ is equipped with zero Dirichlet boundary conditions. Again $W_0(t)$ is nonnegativity-preserving, thus

$$T(x,t)\geqq(W_0(t)(T_0-w))(x)+w(x) \tag{1.18}$$

for all $t\geqq0$. Since $\partial\Omega$ is compact there exists an $x_0\in\partial\Omega$ such that $g(x)\geqq g(x_0)>0$ for all $x\in\partial\Omega$. By the maximum principle $w(x)\geqq g(x_0)$ for all $x\in\Omega$. Meanwhile, if λ_1 is the first eigenvalue of $-\Delta$ then $\lambda_1 > 0$ and there exists a constant K_0 such that for all $h \in C(\bar{\Omega})$

$$\|W_0(t)h\|_\infty \leqq K_0 \|h\|_\infty e^{-\lambda_1 t} \tag{1.19}$$

for all $t \geqq 0$. The result now follows by substituting $T_0 - w$ for h in (1.19); here we can have $\alpha = g(x_0) - \varepsilon$ for any ε with $0 < \varepsilon \ll g(x_0)$.

2. Proof of Theorem 1.2. We first prove the theorem in the case that the $(Z_{i0})_{AV}$ are all positive. Setting $\bar{Z}_i \equiv (1/\alpha_i)\bar{Y}_i - (1/\alpha_1)\bar{Y}_1$ we have, by taking appropriate linear combinations of (1.4b) and integrating over Ω, that $\bar{Z}_i = (Z_{i0})_{AV}$, $i = 2, \ldots, M$, and hence that $\bar{Y}_i = \alpha_i (Z_{i0})_{AV} + (\alpha_i/\alpha_1)\bar{Y}_1$, $i = 2, \ldots, M$. By (1.14) we then have that

$$\lim_{t\to\infty} (Y_i(x, t) - (\alpha_i/\alpha_1)\bar{Y}_1(t)) = \alpha_i (Z_{i0})_{AV}, \qquad i = 2, \ldots, M. \tag{2.1}$$

Then by (2.1) and the nonnegativity of Y_1 we have that there exist constants $b_i > 0$ and $t_i > 0$, $i = 2, \ldots, M$, such that for all $x \in \Omega$ and all $t \geqq t_i$

$$Y_i(x, t) \geqq b_i; \tag{2.2}$$

in (2.2) we can take $b_i = \alpha_i (Z_{i0})_{AV} - \varepsilon$, $i = 2, \ldots, M$, for any ε with $0 < \varepsilon \ll \alpha_i (Z_{i0})_{AV}$. From Proposition 1.2 or 1.3 we have that there exist an $\alpha > 0$ and a $t_1 > 0$ such that

$$T(x, t) \geqq \alpha \tag{2.3}$$

for all $x \in \Omega$ and all $t \geqq t_1$. Set $t_0 = \max\{t_1, t_2, \ldots, t_M\}$; then from (2.2), (2.3), and the monotonicity if f we have that

$$-c\alpha_1 Y_1^{\nu_1}(x, t) \geqq -\alpha_1 \omega(x, t) \tag{2.4}$$

for all $x \in \Omega$ and all $t \geqq t_0$, where $c = b_2^{\nu_2} b_3^{\nu_3}, \ldots, b_M^{\nu_M} f(\alpha)$. Standard comparison principles (see, e.g., [22]) now assert that if $u = u(t)$ is a solution of the ODE initial-value problem

$$\dot{u} = -c\alpha_1 u^{\nu_1}, \tag{2.5a}$$

$$u(0) = \|Y_{10}\|_\infty, \tag{2.5b}$$

then for all $x \in \Omega$ and $t \geqq t_0$

$$0 \leqq Y_1(x, t) \leqq u(t - t_0). \tag{2.6}$$

But (2.5) can be easily solved by the method of separation: if $\nu_1 = 1$ then $u(t) = \|Y_{10}\|_\infty \exp(-c\alpha_1 t)$, while if $\nu_1 > 1$ then $u(t) = (at + b)^{-\gamma_1}$ where $\gamma_1 = 1/(\nu_1 - 1)$, $a = c\alpha_1/(\nu_1 - 1)$, and $b = \|Y_{10}\|_\infty^{1-\nu_1}$. In either case we have from (2.6) that $Y_1(x, t) \to 0$ uniformly in x as $t \to \infty$. Hence $\bar{Y}_1(t) \to 0$ by the bounded convergence theorem. The stated convergence of $Y_2, \ldots, Y_M$ now follows from (2.1). The stated convergence of $T(x, t)$ in the case that conditions (1.6) are assumed follows from (1.11), (1.15), and the convergence of Y_1 (and hence $\bar{Y}_1$) to zero.

We now focus on the case of boundary conditions as in (1.7) and let w be as in (1.8). By replacing T_0 and Y_{i0}, $i = 1, 2, 3$, by $T(x, t_0)$ and $Y_i(x, t_0)$, $i = 1, 2, 3$, if necessary, we can without loss of generality assume that (2.6) holds with $t_0 = 0$. Then, combining (2.6) with (1.19) we have for $K = \|Y_{20}\|_\infty^{\nu_2} \|Y_{30}\|_\infty^{\nu_M} \cdots \|Y_{M0}\|_\infty^{\nu_M} K_0 B$ that

$$\int_0^t W_0(t - s)\omega(s)\, ds \leqq K \int_0^t e^{-\lambda_1 (t-s)} [u(s)]^{\nu_1}\, ds, \tag{2.7}$$

where, on the left-hand side of (2.7), we suppress the x-dependence. In the case $\nu_1 = 1$ we then have, for $\lambda_2 = -c\alpha_1\nu_1$ and $K_1 = K\|Y_{10}\|_\infty$, that

$$\begin{aligned}\int_0^t W_0(t-s)\omega(s)\,ds &\leqq K_1\int_0^t e^{-\lambda_1(t-s)}\,e^{-\lambda_2 s}\,ds\\ &= K_1\,e^{-\lambda_1 t}\int_0^t e^{(\lambda_1-\lambda_2)s}\,ds \qquad (2.8)\\ &= \frac{K_1}{\lambda_1-\lambda_2}(e^{-\lambda_2 t}-e^{-\lambda_1 t}).\end{aligned}$$

We have dealt with the right-hand side of (2.8) in earlier works; see, e.g., [3], [4]. If $\nu_1 > 1$, let γ_1, a, and b be as above so that $u(t) = (at+b)^{-\gamma_1}$, then

$$(2.9)\qquad \int_0^t W_0(t-s)\omega(s)\,ds \leqq K\int_0^t e^{-\lambda_1(t-s)}(as+b)^{-\gamma_2}\,ds,$$

where $\gamma_2 = \gamma_1\nu_1 = \nu_1/(\nu_1-1)$; note that $b > 0$ so that the integral on the right-hand side of (2.9) is well defined.

Since $\gamma_2 > 1$, given $\varepsilon > 0$ we can select a constant $k > 0$ such that

$$(2.10)\qquad K\int_k^\infty (as+b)^{-\gamma_2}\,ds < \frac{\varepsilon}{2}.$$

Then, for this choice of k,

$$\begin{aligned}K\int_0^t e^{-\lambda_1(t-s)}(as+b)^{-\gamma_2}\,ds &\leqq Kb^{-\gamma_2}\,e^{-\lambda_1 t}\int_0^k e^{\lambda_1 s}\,ds+\frac{\varepsilon}{2} \qquad (2.11)\\ &= K_2\,e^{-\lambda_1 t}+\frac{\varepsilon}{2},\end{aligned}$$

where $K_2 = (1/\lambda_1)Kb^{-\gamma_2}(e^{\lambda_1 k}-1)$. We thus see by (2.11) that the right-hand side of (2.9) can be made as small as we wish by choosing t large enough. Hence from (2.8) and (2.9) we see that

$$(2.12)\qquad \lim_{t\to\infty}\int_0^t W_0(t-s)\omega(s)\,ds = 0$$

uniformly in x. The convergence of T to w now follows from (1.17), (2.12), and (1.19) with h replaced by $T_0 - w$.

We now consider the "critical" cases when some or all of the $(Z_{i0})_{AV}$ are zero. By relabeling, if necessary, we can assume that $(Z_{i0})_{AV} = 0$, $i = 2, \dots, M_0$, for some M_0 with $2 \leqq M_0 \leqq M$, and that $(Z_{i0})_{AV} > 0$, $i = M_0+1, \dots, M$. Then from the remarks preceding (2.1) we have that

$$(2.13)\qquad \bar{Y}_i(t) - \left(\frac{\alpha_i}{\alpha_1}\right)\bar{Y}_1(t) = 0, \qquad i = 2, \dots, M_0.$$

Combining (2.13) with (1.14), we obtain that

$$(2.14)\qquad \lim_{t\to\infty}\left\|Y_i(\cdot, t) - \left(\frac{\alpha_i}{\alpha_1}\right)Y_1(\cdot, t)\right\|_\infty = 0, \qquad i = 2, \dots, M_0.$$

Thus given $\varepsilon < 0$ there exists a constant $t_i > 0$ such that

$$(2.15)\qquad Y_i(x, t) \geqq \left(\frac{\alpha_i}{\alpha_1}\right)(Y_1(x, t) - \varepsilon), \qquad i = 2, \dots, M_0$$

for all $x \in \Omega$ and all $t \geqq t_i$. Meanwhile, as in the first part of the proof, we can find constants $b_i > 0$ and $t_i > 0$ such that (2.2) holds for all $x \in \Omega$ and $t \geqq t_i$, $i = M_0 + 1, \ldots, M$. Let t_1 and α also be as in the first part of the proof and set $t_0 = \max\{t_1, \ldots, t_M\}$, $d = \min\{\alpha_2/\alpha_1, \ldots, \alpha_{M_0}/\alpha_1\}$, $\nu_0 = \nu_2 + \cdots + \nu_{M_0}$, and $c = \alpha_1[\prod_{i=M_0+1}^{M} b^{\nu_i}]d^{\nu_0}f(\alpha)$. Then standard comparison principles now assert that if $u = u(t)$ solves the ODE initial-value problem

$$\dot{u} = -cu^{\nu_1}(u - \varepsilon)^{\nu_0} \tag{2.16a}$$

$$u(0) = \| Y_{10} \|_\infty, \tag{2.16b}$$

then for all $x \in \Omega$ and all $t \geqq t_0$

$$0 \leqq Y_1(x, t) \leqq u(t - t_0). \tag{2.17}$$

Setting $u_\varepsilon(t) = u(t) - \varepsilon$, we can rewrite (2.16a) as

$$\dot{u}_\varepsilon = -c(u_\varepsilon + \varepsilon)^{\nu_1}(u_\varepsilon)^{\nu_0}. \tag{2.18}$$

Then $u_\varepsilon(t) \leqq \bar{u}_\varepsilon(t)$ where $\bar{u}_\varepsilon$ solves the ODE initial-value problem

$$\dot{\bar{u}}_\varepsilon = -c(\bar{u}_\varepsilon)^{\nu} \tag{2.19a}$$

$$\bar{u}_\varepsilon(0) = \| Y_{10} \|_\infty - \varepsilon \tag{2.19b}$$

with $\nu = \nu_1 + \nu_0$. Since $\nu_i \geqq 1$, $i = 1, \ldots, M$, we have that $\nu > 1$; hence $\bar{u}_\varepsilon(t) = (at + b)^{-\gamma}$ where $\gamma = 1/(\nu - 1)$, $a = c/(\nu - 1)$, and $b = (\| Y_{10} \|_\infty - \varepsilon)^{1-\nu}$. Since $u(t) = u_\varepsilon(t) + \varepsilon \leqq \bar{u}_\varepsilon(t) + \varepsilon$, we have from (2.17) that for all $x \in \Omega$ and all $t \geqq t_0$

$$0 \leqq Y_1(x, t) \leqq (a(t - t_0) + b)^{-\gamma} + \varepsilon. \tag{2.20}$$

Since ε is arbitrary we conclude from (2.14) and (2.20) that $Y_i(x, t) \to 0$ uniformly, $i = 1, \ldots, M_0$, as $t \to \infty$.

The stated convergence of T in the case that conditions (1.6) are assumed now follows exactly as before from (1.11). In the case that conditions (1.7) are assumed, we need only slightly modify our arguments: (2.9) becomes

$$\begin{aligned} \int_0^t W_0(t - s)\omega(s)\,ds &\leqq K \int_0^t e^{-\lambda_1(t-s)}(as + b)^{-\gamma}\,ds + \varepsilon K \int_0^t e^{-\lambda_1(t-s)}\,ds \\ &\leqq K \int_0^t e^{-\lambda_1(t-s)}(as + b)^{-\gamma}\,ds + \varepsilon\left(\frac{1}{\lambda_1}\right)K. \end{aligned} \tag{2.21}$$

The integral on the right-hand side of (2.21) goes to zero exactly as before; combining this with the fact that ε is arbitrary we see that (2.12) again holds here. The convergence of T to w now follows as before, thus completing the proof of the theorem.

3. Proof of Lemma 1.1. We have seen in § 1 that T and the $Y_i(x, t)$ are uniformly bounded (specifically $\| Y_i(\cdot, t) \|_\infty \leqq \| Y_{i0} \|_\infty$, $i = 1, \ldots, M$, and $\| T(\cdot, t) \|_\infty \leqq M_T$ for all $t \geqq 0$). Hence by the Schauder estimates (see, e.g., [12]) there are uniform bounds for $|\nabla T|$ and $|\nabla Y_i|$ as well, $i = 1, \ldots, M$. Differentiating (1.4) through by ∇, and using these gradient estimates, the product and chain rules, and the smoothness of f, we then obtain uniform bounds for $|\partial/\partial t \nabla T|$ and $|\partial/\partial t \nabla Y_i|$, $i = 1, \ldots, M$. Let $Q_t = \Omega \times [0, t]$ for each $t > 0$. Integrating (1.4b) over Q_t we obtain

$$\| Y_1(\cdot, t) \|_1 + \alpha_1 \int_0^t \int_\Omega \omega(x, s)\,dx\,ds \leqq \| Y_{10} \|_1 \tag{3.1}$$

where $\|\cdot\|_p$ denotes the norm on $L^p(\Omega)$, $1 \leqq p < \infty$. Hence

$$\int_0^\infty \int_\Omega \omega(x, s)\, dx\, ds \leqq \left(\frac{1}{\alpha_1}\right) \| Y_{10} \|_1. \tag{3.2}$$

Multiplying (1.4b) by Y_i and integrating over Q_t we obtain

$$\int_0^t \int_\Omega |\nabla Y_i|^2\, dx\, ds \leqq \frac{1}{2} \| Y_{i0} \|_2^2, \qquad i = 1, \dots, M. \tag{3.3}$$

If conditions (1.6) are assumed, we multiply (1.4a) by T and integrate over Q_t to obtain

$$\frac{1}{2} \| T(\cdot, t) \|_2^2 + \int_0^t \int_\Omega |\nabla T|^2\, dx\, ds \leqq QM_T \int_0^t \int_\Omega \omega(x, s)\, dx\, ds + \frac{1}{2} \| T_0 \|_2^2. \tag{3.4}$$

From (3.3) and by combining (3.2) and (3.4) we thus see that

$$\int_0^\infty \int_\Omega |\nabla Y_i|^2\, dx\, ds < \infty, \qquad i = 1, \dots, M, \tag{3.5}$$

and if conditions (1.6) are assumed that

$$\int_0^\infty \int_\Omega |\nabla T|^2\, dx\, ds < \infty. \tag{3.6}$$

The uniform bounds on $|\partial/\partial t \nabla T|$ and $|\partial/\partial t \nabla Y_i|$, $i = 1, \dots, M$, Hölder's inequality, and (3.5), (3.6) then show that

$$\lim_{t\to\infty} \| |\nabla Y_i(\cdot, t)| \|_p = 0, \qquad i = 1, \dots, M \tag{3.7}$$

for all $p \geqq 1$ and, if conditions (1.6) are assumed, that

$$\lim_{t\to\infty} \| |\nabla T(\cdot, t)| \|_p = 0 \tag{3.8}$$

for all $p \geqq 1$.

If λ_i is the first positive eigenvalue of the operator $-d_i\Delta$ equipped with zero Neumann boundary conditions, then by eigenfunction expansion we see that

$$\lambda_i \| Y_i(\cdot, t) - \bar{Y}_i(t) \|_2 \leqq \| |\nabla Y_i(\cdot, t)| \|_2^2, \qquad i = 1, \dots, M \tag{3.9}$$

and hence

$$\lim_{t\to\infty} \| Y_i(\cdot, t) - \bar{Y}_i(t) \|_2 = 0, \qquad i = 1, \dots, M. \tag{3.10}$$

From (3.10), the uniform bounds on the Y_i, and Hölder's inequality, we then have that

$$\lim_{t\to\infty} \| Y_i(\cdot, t) - \bar{Y}_i(t) \|_p = 0, \qquad i = 1, \dots, M \tag{3.11}$$

for all $p \geqq 1$. Selecting $p > n$, we have from the standard Sobolev embedding theorems that there exists a constant $K = K(n, p, \Omega)$ such that

$$\| Y_i(\cdot, t) - \bar{Y}_i(t) \|_\infty \leqq K[\| Y_i(\cdot, t) - \bar{Y}_i(t) \|_p + \| |\nabla Y_i(\cdot, t)| \|_p], \qquad i = 1, \dots, M, \tag{3.12}$$

where we note that $\nabla \bar{Y}_i(t) = 0$. Thus (1.14) follows from (3.7), (3.11), and (3.12). Similarly, starting with (3.8), (1.15) follows if conditions (1.6) are assumed. This completes the proof of the lemma.

Remark. From (3.6) on, the arguments used here are virtually identical to those employed in the proof of [8, Lemma 2.5].

REFERENCES

[1] J. D. AVRIN, *Qualitative theory for a model of laminar flames with arbitrary nonnegative initial data*, J. Differential Equations, 84 (1990), pp. 290-308.

[2] ———, *Qualitative theory of the Cauchy problem for a one-step reaction model on bounded domains*, SIAM J. Math. Anal., 22 (1991), pp. 379-391.

[3] ———, *Decay and boundedness results for a model of laminar flames with complex chemistry*, Proc. Amer. Math. Soc., 110 (1990), pp. 989-995.

[4] ———, *Flame propagation versus extinction in models of complex chemistry*, submitted.

[5] 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.

[6] J. BUCKMASTER AND G. S. S. LUDFORD, *Theory of Laminar Flames*, Cambridge University Press, Cambridge, 1982.

[7] P. CLAVIN, *Dynamical behavior of premixed fronts in laminar and turbulent flows*, Prog. Energy. Comb. Sci., 11 (1985), pp. 1-59.

[8] W. E. FITZGIBBON AND C. B. MARTIN, *Semilinear parabolic systems modelling spatially inhomogeneous exothermic reactions*, preprint.

[9] I. M. GELFAND, *Some problems in the theory of quasilinear equations*, Amer. Math. Soc. Transl. Ser. 2, 29 (1963), pp. 295-381.

[10] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, New York, 1981.

[11] 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.

[12] O. A. LADYZENSKAJA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and quasilinear equations of parabolic type*, Trans. Amer. Math. Soc., 23, Providence, RI, 1968.

[13] B. LARROUTUROU, *The equations of one-dimensional unsteady flame propagation: existence and uniqueness*, SIAM J. Math. Anal., 19 (1988), pp. 32-59.

[14] M. MARION, *Etude mathématique d'un modèle de flamme laminaire sans température d'ignition: I-Cas scalaire*, Ann. Fac. Sci. Toulouse Math., 6 (1984), pp. 215-244.

[15] ———, *Qualitative properties of a nonlinear system for laminar flames without ignition temperature*, Nonlinear Anal. TMA, 9 (1985), pp. 1269-1292.

[16] ———, *Attractors for reaction-diffusion equations: existence and estimate of their dimension*, Appl. Anal., 25 (1987), pp. 101-147.

[17] V. H. MOLL AND A. I. FOGELSON, *Activation waves in a model of platelet aggregation: existence of solutions and stability of traveling fronts*, J. Math. Biol., to appear.

[18] V. H. MOLL, *Threshold phenomena in a model for platelet aggregation: existence of global solutions and critical multipliers*, Nonlinear Anal. TMA, to appear.

[19] J. M. ROQUEJOFFRE, Thesis, INRIA Sophia Antipolis, June 1988.

[20] D. SATTINGER, *A nonlinear parabolic system in the theory of combustion*, Quart. Appl. Math., 33 (1975), pp. 47-61.

[21] G. I. SIVASHINSKY, *Instabilities, pattern formation, and turbulence in flames*, Ann. Rev. Fluid Mech., 15 (1988), pp. 179-199.

[22] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

[23] D. TERMAN, *Connection problems arising from nonlinear diffusion equations*, in the Proceedings of the Microconference on Nonlinear Diffusion, J. Serrin, L. Peletier, W.-M. Ni, eds., Springer-Verlag, Berlin-Heidelberg-New York, 1986.

[24] ———, *Stability of planar wave solutions to a combustion model*, SIAM J. Math. Anal., 21 (1990), pp. 1139-1171.

[25] D. H. WAGNER, *Premixed laminar flames as travelling waves*, in Reacting Flows: Combustion and Chemical Reactors, G. S. S. Ludford, ed., AMS Lectures in Applied Mathematics, 24, Providence, RI, 1986.

[26] ———, *Existence of deflagration waves: connection to a degenerate critical point*, Proceedings of the Conference on Physical Mathematics and Nonlinear Partial Differential Equations, J. H. Lightbourne and S. M. Rankin, eds., West Virginia University, Marcel-Dekker, New York, 1985.

[27] F. WILLIAMS, *Combustion Theory*, Second ed., Addison-Wesley, Reading, MA, 1985.

SIAM J. MATH. ANAL.
Vol. 24, No. 2, pp. 299–316, March 1993

003

ON A DISSOLUTION-GROWTH PROBLEM WITH SURFACE TENSION IN THE NEIGHBORHOOD OF A STATIONARY SOLUTION*

F. ABERGEL[†], D. HILHORST[†], AND F. ISSARD-ROCH[†]

Abstract. The authors consider a one-phase Stefan problem with surface tension in dimension two and show a well-posedness result in the neighborhood of a stationary solution, in the case that the moving interface is parametrized in the form $y = f(x,t)$.

Key words. Stefan problem, surface tension, free boundary, nonlinear parabolic PDEs

AMS(MOS) subject classifications. 35K55, 35R35

1. Introduction.

Consider a system composed of a solid phase of a single compound and an incompressible liquid phase which is a dilute solution of that compound, and suppose that the evolution in time of the system is governed by two processes: a diffusion process in a diffusion layer in the fluid and a dissolution-growth process, located at the interface between solid and fluid.

The mathematical problem can be described by three basic equations

$$C_t = \Delta C \quad \text{in the diffusion layer,} \tag{1.1}$$

where $C = C(x,y,t)$ is the concentration in the liquid phase,

$$\frac{\partial C}{\partial n}|_{\Gamma(t)} = V_n, \tag{1.2}$$

where $\Gamma(t)$ denotes the interface between solid and fluid, n is the normal unit vector to the interface directed towards the fluid, and V_n is the normal velocity of $\Gamma(t)$,

$$V_n = -\alpha e^{\gamma K} + C|_{\Gamma(t)} \tag{1.3}$$

where α is a positive constant, the positive constant γ is proportional to the surface tension of the interface, and K is the curvature of the interface. Let us note that in view of (1.3), the boundary condition (1.2) can be rewritten as

$$\left(C - \frac{\partial C}{\partial n}\right)|_{\Gamma(t)} = \alpha e^{\gamma K}.$$

In this paper, we consider the case where the interface can be parametrized in the form $y = f(x,t)$ and denote by e the width of the diffusion layer. Then f and C

*Received by the editors August 13, 1991; accepted for publication (in revised form) July 9, 1992. Part of this project was finished while two of the authors (F.A. and D.H.) were visiting the Institute for Mathematics and Its Applications at the University of Minnesota.

[†]Laboratoire d' Analyse Numérique, Centre National de la Recherche Scientifique, et Université Paris Sud, 91405 Orsay Cédex, France.

satisfy the following (rescaled) problem:

$$P_0 \begin{cases} C_t = \Delta C \quad \text{in } Q_f := \{(x,y,t) \in \mathbb{R}^2 \times \mathbb{R}^+, f(x,t) < y < f(x,t)_+e\}, & \\ \left(C - \frac{\partial C}{\partial n}\right)(x, f(x,t), t) = \alpha e^{-\gamma(f_{xx}/(1+f_x^2)^{3/2})(x,t)}, & (x,t) \in \mathbb{R} \times \mathbb{R}^+, \\ \frac{\partial C}{\partial n}(x, f(x,t)+e, t) = 0, & (x,t) \in \mathbb{R} \times \mathbb{R}^+, \\ C(x+2L, y, t) = C(x,y,t), & (x,y,t) \in Q_f, \\ \frac{f_t}{(1+f_x^2)^{1/2}} = -\alpha e^{-\gamma f_{xx}/(1+f_x^2)^{3/2}} + C(x, f(x,t), t), & (x,t) \in \mathbb{R} \times \mathbb{R}^+, \\ f(x+2L, t) = f(x,t), & (x,t) \in \mathbb{R} \times \mathbb{R}^+, \\ f(x,0) = f_0(x), & x \in \mathbb{R}, \\ C(x,y,0) = C_0(x,y) \quad (x,y) \in \Sigma_f := \{(x,y) \in \mathbb{R}^2, f_0(x) < y < f_0(x)_+e\}, & \end{cases}$$

where f_0 and C_0 satisfy the hypothesis H_0:

$f_0 \in C_{\text{per}}^{3+\lambda}(\mathbb{R}), C_0 \in C_{\text{per}}^{2+\lambda}(\overline{\Sigma}_f)$ for some $\lambda \in (0,1)$, and C_0 satisfies the compatibility conditions

$$\frac{\partial C_0}{\partial n}(x, f_0(x)+e) = 0 \qquad x \in \mathbb{R},$$
$$\left(C_0 - \frac{\partial C_0}{\partial n}\right)(x, f_0(x)) = \alpha e^{-\gamma(f_0''/(1+f_0'^2)^{3/2})}(x) \qquad x \in \mathbb{R}.$$

(By the subscript per we denote functions periodic in the variable x with period $2L$.)

We remark that pairs of the form $(C, f) = (\alpha,$ constant $)$ are stationary solutions of problem P_0. The purpose of this paper is to prove the following existence and uniqueness result.

THEOREM 1.1. *There exists a positive constant $\rho_0 = \rho_0(T)$ such that if the initial data satisfies the condition*

$$\|f_0\|_{\mathbb{R}}^{(3+\lambda)} + \|C_0 - \alpha\|_{\Sigma_f}^{(2+\lambda)} \leq \rho_0,$$

then P_0 has a unique solution (f, C) in $C_{\text{per}}^{3+\lambda,(3+\lambda)/2}(\mathbb{R} \times [0,T]) \times C_{\text{per}}^{2+\lambda,1+\lambda/2}(\overline{Q_f^T})$, where $Q_f^T := \{(x,y,t) \in Q_f, t \in (0,T)\}$.

The paper is organized as follows. In §2, we present a physical derivation of the problem which is due to Cournil [5]. In §3, we transform problem P_0 into a problem on a fixed domain. In §4, we recall the definitions of the Hölder spaces that we use and prove some properties which are useful in the following sections. In §5, we consider a related linear problem and prove its well-posedness.

We prove Theorem 1.1 transposed to the coordinates of the fixed domain in §§6 and 7. An improved uniqueness result is given in §6. The existence proof is given in §7. It is done by means of a fixed point method: we iteratively solve the linear problem of §5, while considering as known the nonlinear terms. The idea is to show that the iterative map is a strict contraction from a ball of small enough radius $\rho > \rho_0$ into itself.

Let us remark that our proof would carry over to the case of a domain given by

$$\{(x,y,t) \in \mathbb{R}^2 \times \mathbb{R}^+, f(x,t) < y < A\}$$

with A a fixed constant. A suitable change of variables for transforming the problem to a fixed domain is then given by $\tilde{y} = (A-y)/(A-f(x,t))$, and the constants ρ_0

and ρ should be chosen small enough so that the interface does not hit the boundary $y = A$.

In a forthcoming paper we shall extend our method to the proof of local existence and uniqueness for arbitrary initial data. For the well-posedness of a very similar problem in the case of spherical symmetry we refer to [4], [7].

For local existence and uniqueness results for related problems we refer to a study of Duchon and Robert [6] of a quasistationary problem and to papers by X. Y. Chen [3] and Xinfu Chen [2] about a two phase problem.

2. The physical derivation. In heterogeneous media, a problem of interest is the creation or the consumption of one or more phases. These phenomena are accompanied by a change in the geometry of the reaction field and, in particular, by an evolution of interfaces. The interfaces are of essential importance: on the one hand, they determine the geometrical configuration of the system and on the other hand they intervene, by means of their curvature, in the equilibrium and stability conditions.

In this paper, we consider a system composed of a solid phase of a single compound and an incompressible liquid phase, which is a dilute solution of that compound. The evolution in time of this system induces mass transfer processes: a homogeneous one which consists of a diffusion process in a diffusion layer in the fluid and a heterogeneous one, namely, a dissolution-growth process, located at the interface between solid and liquid.

Let $y = f(x,t)$ be a parametrization of the interface, let e denote the width of the diffusion layer, and let $C(x,y,t)$ with $f(x,t) < y < f(x,t) + e$ denote the concentration in the liquid phase. The equations of the problem are deduced from the following physical laws.

(i) *Mass balance.* We shall assume that the fluid is essentially at rest, so that the convective velocity is negligible; let J denote the diffusion flux, which we suppose given by Fick's law:

$$J = -D \operatorname{grad} C, \tag{2.1}$$

where the diffusion coefficient D is a positive constant.

Upon writing the conservation of mass for an arbitrary subdomain ω consisting of the same particles for each time, we obtain

$$0 = \frac{d}{dt}\left(\int_{\omega_t} C dx\right) = \int_{\omega_t} C_t dx + \int_{\partial\omega_t^l} J.\vec{n} d\sigma + \int_{\partial\omega_t^s} (C - V^{-1})\vec{v}.\vec{n} d\sigma. \tag{2.2}$$

In (2.2) $\partial\omega_t^l$ (respectively, $\partial\omega_t^s$) is that part of $\partial\omega_t$ which makes contact with the liquid (respectively, solid) phase, V^{-1} is the concentration in the solid phase, and $\vec{v}$ is the velocity of the interface.

From (2.2) we easily obtain the governing equations

$$C_t = D\Delta C \tag{2.3}$$

in the fluid domain, and

$$D\frac{\partial C}{\partial n} = (V^{-1} - C)\vec{v}.\vec{n} \tag{2.4}$$

at the interface.

From now on, we shall assume that V^{-1} is much larger than C, and write (2.4) as

$$D\frac{\partial C}{\partial n} = V^{-1}\vec{v}.\vec{n}. \tag{2.5}$$

Upon using the fact that the interface is the graph of a function $f(x,t)$, we eventually write (2.5) under its final form

$$D\frac{\partial C}{\partial n} = V^{-1}\frac{ft}{\sqrt{1+f_x^2}}. \tag{2.6}$$

(ii) *Dissolution and growth of the solid.* We suppose that the kinetics of dissolution and growth follows Nernst's law, which derives, either from a rate determining diffusion in a layer of constant thickness, or from a rate determining first order interface reaction [8], namely,

$$J\cdot n|_{y=f(x,t)} = -h(C,R), \tag{2.7}$$

where R is the radius of curvature of the interface

$$R(x,t) = -\frac{(1+f_x^2)^{3/2}}{f_{xx}} \tag{2.8}$$

and the function h is given by

$$h(C,R) = K(C - s_0 e^{\gamma/R}), \tag{2.9}$$

where K is a kinetic constant, s_0 is the concentration at saturation of the solution, and γ is proportional to the surface tension of the interface.

Using (2.6)–(2.9), we deduce that

$$\frac{1}{V}f_t = (1+f_x^2)^{1/2}K(C(x,f(x,t),t) - s_0 e^{-\gamma f_{xx}/(1+f_x^2)^{3/2}}), \tag{2.10}$$

whereas (2.1) and (2.7)–(2.9) imply the boundary condition

$$D\frac{\partial C}{\partial n} = K(C - s_0 e^{-\gamma f_{xx}/(1+f_x^2)^{3/2}}) \tag{2.11}$$

on the interface $y = f(x,t)$, where n is the normal to the interface directed towards the fluid.

(iii) *Other boundary conditions.* We assume that C and f are periodic in the x-direction, namely,

$$\begin{cases} C(x+2L,y,t) = C(x,y,t) & \text{in } Q_f, \\ f(x+2L,t) = f(x,t) & \text{in } \mathbb{R}\times\mathbb{R}_+, \end{cases}$$

and that C satisfies the Neumann condition

$$\frac{\partial C}{\partial n}(x,f(x,t)+e,t) = 0, \qquad (x,t)\in\mathbb{R}\times\mathbb{R}^+. \tag{2.12}$$

Finally we set

$$\tilde{x} = \frac{Kx}{D}, \quad \tilde{y} = \frac{Ky}{D}, \quad \tilde{t} = \frac{K^2 t}{D}, \quad \tilde{C} = VC, \quad \tilde{\gamma} = \frac{K\gamma}{D},$$
$$\tilde{f} = \frac{Kf}{D}, \quad \alpha = Vs_0, \quad \tilde{L} = \frac{KL}{D}, \quad \tilde{e} = \frac{Ke}{D}$$

and omit the tilde. The equations (2.3), (2.11), (2.12), and (2.10) transform into

$$\begin{aligned}
&C_t = \Delta C \quad \text{in } Q_f := \{(x,y,t) \in \mathbb{R}^2 \times \mathbb{R}^+, f(x,t) < y < f(x,t) + e\},\\
&\left(C - \frac{\partial C}{\partial n}\right)(x, f(x,t), t) = \alpha e^{-\gamma(f_{xx}/(1+f_x^2)^{3/2})(x,t)}, \qquad (x,t) \in \mathbb{R} \times \mathbb{R}^+,\\
&\frac{\partial C}{\partial n}(x, f(x,t) + e, t) = 0, \qquad (x,t) \in \mathbb{R} \times \mathbb{R}^+,\\
&f_t = (1 + f_x^2)^{1/2}\left(-\alpha e^{-\gamma f_{xx}/(1+f_x^2)^{3/2}} + C(x, f(x,t), t)\right), \qquad (x,t) \in \mathbb{R} \times \mathbb{R}^+,
\end{aligned}$$

which, together with periodic boundary conditions and periodic initial conditions for C and f, yield Problem P_0.

3. Transformation of the problem to a problem on a fixed domain. Let us define new coordinates by

$$\hat{x} = x, \quad \hat{y} = y - f(x,t), \quad \hat{t} = t$$

and set

$$g(x,y,t) = C(x,y,t) - \alpha, \quad \hat{g}(\hat{x},\hat{y},\hat{t}) = g(x,y,t), \quad \hat{f}(\hat{x},\hat{t}) = f(x,t).$$

Some easy computations permit us to show that $\hat{g}$ and $\hat{f}$ satisfy

$$P\left\{\begin{array}{l}
\hat{g}_{\hat{t}} = \hat{g}_{\hat{x}\hat{x}} + \hat{g}_{\hat{y}\hat{y}} + \hat{f}_{\hat{x}}^2 \hat{g}_{\hat{y}\hat{y}} - (\hat{f}_{\hat{x}\hat{x}} - \hat{f}_{\hat{t}})\hat{g}_{\hat{y}} - 2\hat{f}_{\hat{x}}\hat{g}_{\hat{x}\hat{y}}\\
\qquad\qquad \text{in } Q_0 = \{(x,y,t) \in \mathbb{R}^2 \times \mathbb{R}^+, 0 < y < e\},\\
\hat{g}(\hat{x},0,\hat{t}) - (1 + \hat{f}_{\hat{x}}^2)^{1/2}\hat{g}_{\hat{y}}(\hat{x},0,\hat{t}) + \frac{\hat{f}_{\hat{x}}}{(1+\hat{f}_{\hat{x}}^2)^{1/2}}\hat{g}_{\hat{x}}(\hat{x},0,\hat{t})\\
\quad = \alpha(e^{-\gamma \hat{f}_{\hat{x}\hat{x}}/(1+\hat{f}_{\hat{x}}^2)^{3/2}} - 1) \quad (\hat{x},\hat{t}) \in \mathbb{R} \times \mathbb{R}^+\\
(1 + \hat{f}_{\hat{x}}^2)^{1/2}\hat{g}_{\hat{y}}(\hat{x},e,\hat{t}) - \frac{\hat{f}_x}{(1+\hat{f}_{\hat{x}}^2)^{1/2}}\hat{g}_{\hat{x}}(\hat{x},e,\hat{t}) = 0 \quad (\hat{x},\hat{t}) \in \mathbb{R} \times \mathbb{R}^+\\
\hat{g}(\hat{x} + 2L, \hat{y}, \hat{t}) = \hat{g}(\hat{x},\hat{y},\hat{t}) \quad \text{in } Q_0\\
\hat{g}(\hat{x},\hat{y},0) = \hat{g}_0(\hat{x},\hat{y}) \quad (\hat{x},\hat{y}) \in \mathbb{R} \times (0,e),\\
\text{where}\\
\hat{g}_0(\hat{x},\hat{y}) = C_0(x,y) - \alpha \quad (\text{with } \hat{y} = y - f_0(x)),\\
\text{and}\\
\hat{f}_{\hat{t}} = -(1 + \hat{f}_{\hat{x}}^2)^{1/2}(\alpha e^{-\gamma \hat{f}_{\hat{x}\hat{x}}/(1+\hat{f}_{\hat{x}}^2)^{3/2}} - \hat{g}(\hat{x},0,\hat{t}) - \alpha) \quad (\hat{x},\hat{t}) \in \mathbb{R} \times \mathbb{R}^+\\
\hat{f}(\hat{x} + 2L, \hat{t}) = \hat{f}(\hat{x},\hat{t}) \quad (\hat{x},\hat{t}) \in \mathbb{R} \times \mathbb{R}^+\\
\hat{f}(\hat{x},0) = \hat{f}_0(\hat{x}) \quad \hat{x} \in \mathbb{R},\\
\text{where}\\
\hat{f}_0(\hat{x}) = f_0(x).
\end{array}\right.$$

Remark 3.1. In what follows, it will be convenient to write

$$\begin{aligned}(1+f_x^2)^{1/2}&(e^{-\gamma f_{xx}/(1+f_x^2)^{3/2}}-1)\\&=-\gamma f_{xx}+(1+f_x^2)^{1/2}(e^{-\gamma f_{xx}/(1+f_x^2)^{3/2}}-1+\gamma f_{xx}/(1+f_x^2)^{1/2})\end{aligned}$$

and

$$(1+f_x^2)^{1/2}g(x,0,t)=g(x,0,t)-(1-(1+f_x^2)^{1/2})g(x,0,t);$$

this permits us to separate linear from nonlinear terms.

4. A few properties of Hölder spaces. In this paper we work with the Hölder spaces $C^{\ell,\ell/2}(\overline{\Omega}\times[0,T])$ as defined by Ladyženskaja, Solonnikov, and Ural'ceva [9]. We indicate below some notations and properties which will be useful in what follows.

Let $\Omega=\mathbb{R}\times\Omega'$, where Ω' is a smooth open set of $\mathbb{R}^{N-1}$ and $Q_T=\Omega\times(0,T)$. We define the space

$$C_{\text{per}}^{\ell,\ell/2}(\overline{Q}_T)=\Big\{f\in C^{\ell,\ell/2}(\overline{Q}_T),\\ f(x_1+2L,x',t)=f(x_1,x',t)\text{ for all }(x_1,x',t)\in\mathbb{R}\times\Omega'\times(0,T)\Big\}$$

of $2L$-periodic functions in the x_1-direction and define $C_{\text{per}}^{\ell}(\overline{\Omega})$ in a similar way.

In what follows we often use the notation

$$C^{m+\lambda,(m+\lambda)/2}(\overline{Q}_T);\quad\text{then } m=[\ell]\quad\text{and}\quad\lambda=\ell-[\ell].$$

Next we state some technical results which will be useful in the sequel.

LEMMA 4.1. (i) *Let $m\in\mathbb{N}$. There exists a constant $C>0$ which depends only on m such that*

$$\|fg\|_{Q_T}^{(m+\lambda)}\le C\|f\|_{Q_T}^{(m+\lambda)}\|g\|_{Q_T}^{(m+\lambda)}$$

for all $f,g\in C^{m+\lambda,(m+\lambda)/2}(\overline{Q}_T), 0<\lambda<1$.

(ii) *There exists a constant $C>0$ such that*

$$\left\|\frac{1}{f}\right\|_{Q_T}^{(1+\lambda)}\le C(\|f\|_{Q_T}^{(1+\lambda)})^2$$

for all $f\in C^{1+\lambda,(1+\lambda)/2}(\overline{Q}_T)$ such that $f\ge 1$.

(iii) *There exists a constant $C>0$ such that*

$$\|\varphi_o f\|_{Q_T}^{(1+\lambda)}\le C\|\varphi\|_{(-M,M)}^{(1+1)}\max\Big\{1,(\|f\|_{Q_T}^{(1+\lambda)})^2\Big\}$$

for all $f\in C^{1+\lambda,(1+\lambda)/2}(\overline{Q}_T)$ such that $-M\le f\le M$ and all $\varphi\in C^{1+1}([-M,M])$ (by which we mean the space of continuously differentiable functions with Lipschitz continuous derivatives).

Remark. We will use Lemma 4.1 (i) in cases where $g\in C^{m+\lambda,(m+\lambda)/2}(\overline{Q}_T)$ and $f\in C^{m+\lambda,(m+\lambda)/2}(\overline{\Sigma}_T)$ with $\Sigma_T=\mathbb{R}\times(0,T)$. Then we implicitly work with $\tilde{f}(x,y,t)=f(x,t)$ for all $(x,y,t)\in\overline{Q}_T$ so that $\tilde{f}\in C^{m+\lambda,(m+\lambda)/2}(\overline{Q}_T)$ and $\|\tilde{f}\|_{Q_T}^{(m+\lambda)}=\|f\|_{\Sigma_T}^{(m+\lambda)}$.

In what follows we will have to estimate terms of the form

$$E = \exp\left(-\gamma \frac{f_{xx}}{(1+f_x^2)^{3/2}}\right) - 1 + \frac{\gamma f_{xx}}{(1+f_x^2)^{1/2}}$$

as well as differences of such terms. To that purpose we set

$$\chi(s) = e^{\gamma s} - 1 - \gamma s.$$

Then

$$E = \chi\left(-\frac{f_{xx}}{(1+f_x^2)^{3/2}}\right) + \gamma f_{xx}\left(\frac{1}{(1+f_x^2)^{1/2}} - \frac{1}{(1+f_x^2)^{3/2}}\right). \tag{4.1}$$

Next we prove the following result.

Lemma 4.2. *There exists* $\delta_0 > 0$ *such that for all* $\delta \in (0, \delta_0)$ *and for all* $r \in C^{1+\lambda,(1+\lambda)/2}\,(\overline{\Sigma}_T)$ *satisfying* $\|r\|^{(1+\lambda)}_{\Sigma_T} \leq \delta$ *there holds*

$$\|\chi(r)\|^{(1+\lambda)}_{\Sigma_T} \leq \gamma^2 C(\|r\|^{(1+\lambda)}_{\Sigma_T})^2,$$

where C *is the constant introduced in Lemma* 4.1 (i).

Proof. We use the Taylor expansion

$$\chi(s) = \sum_{k=2}^{+\infty} \frac{\gamma^k s^k}{k!} = s^2 \sum_{k=0}^{\infty} \frac{\gamma^{k+2} s^k}{(k+2)!};$$

then we deduce from Lemma 4.1 (i) that, with C being the constant introduced there,

$$\begin{aligned}
\|\chi(r)\|^{(1+\lambda)}_{\Sigma_T} &\leq C\|r^2\|^{(1+\lambda)}_{\Sigma_T} \sum_{k=0}^{\infty} \frac{\gamma^{k+2}\|r^k\|^{(1+\lambda)}_{\Sigma_T}}{(k+2)!} \\
&\leq C^2(\|r\|^{(1+\lambda)}_{\Sigma_T})^2 \left\{ \sum_{k=0}^{\infty} \frac{\gamma^{k+2}}{(k+2)!} C^{k-1} (\|r\|^{1+\lambda}_{\Sigma_T})^k \right\} \\
&\leq \frac{1}{C\delta^2}(\|r\|^{(1+\lambda)}_{\Sigma_T})^2 \sum_{k=0}^{\infty} \frac{(\gamma C\delta)^{k+2}}{(k+2)!} \\
&\leq \frac{1}{C\delta^2}(e^{\gamma C\delta} - 1 - \gamma C\delta)(\|r\|^{(1+\lambda)}_{\Sigma_T})^2.
\end{aligned}$$

Lemma 4.3. *There exists* $\delta_0 > 0$ *such that for all* $\delta \in (0, \delta_0)$ *and for all* $r, r' \in C^{1+\lambda,(1+\lambda)/2}(\overline{\Sigma}_T)$ *satisfying* $\|r\|^{(1+\lambda)}_{\Sigma_T}, \|r'\|^{(1+\lambda)}_{\Sigma_T} \leq \delta$ *there exists* $K = K(\delta_0)$ *such that*

$$\|\chi(r) - \chi(r')\|^{(1+\lambda)}_{\Sigma_T} \leq K\|r - r'\|^{(1+\lambda)}_{\Sigma_T} \max\left(\|r\|^{(1+\lambda)}_{\Sigma_T}, \|r'\|^{(1+\lambda)}_{\Sigma_T}\right).$$

Proof. We have that

$$\begin{aligned}
\chi(s) - \chi(s') &= \sum_{k=2}^{\infty} \frac{\gamma^k}{k!}(s^k - s'^k) \\
&= (s - s') \sum_{k=2}^{\infty} \frac{\gamma^k}{k!} \sum_{p=0}^{k-1} s^p s'^{k-1-p},
\end{aligned}$$

which yields

$$\begin{aligned}\|\chi(r)-\chi(r')\|_{\Sigma_T}^{(1+\lambda)} &\le C\|r-r'\|_{\Sigma_T}^{(1+\lambda)}\sum_{k=2}^{\infty}\frac{\gamma^k}{k!}kC^{k-2}\delta^{k-1}\\ &\le \|r-r'\|_{\Sigma_T}^{(1+\lambda)}\sum_{j=1}^{\infty}\gamma\frac{\gamma^j C^j\delta^j}{j!}\\ &\le \gamma(e^{\gamma C\delta}-1)\|r-r'\|_{\Sigma_T}^{(1+\lambda)}\\ &\le \widetilde{K}\|r-r'\|_{\Sigma_T}^{(1+\lambda)}\max\left(\|r\|_{\Sigma_T}^{(1+\lambda)},\|r'\|_{\Sigma_T}^{(1+\lambda)}\right),\end{aligned}$$

where

$$\widetilde{K}=\gamma\frac{e^{\gamma C\delta}-1}{\delta}\le\gamma^2Ce^{\gamma C\delta_0}.$$

5. The linear problem. In this section we prove that there exists a unique solution of the following linear problem:

$$g_t-\Delta g=F_1 \quad \text{in } Q_T:=\{(x,y,t)\in\mathbb{R}\times(0,e)\times(0,T)\}, \tag{5.1}$$

$$(g-g_y)(x,0,t)+\alpha\gamma f_{xx}(x,t)=F_2(x,t),\ (x,t)\in\Sigma_T:=\mathbb{R}\times(0,T), \tag{5.2}$$

$$g_y(x,e,t)=F_3(x,t),\qquad (x,t)\in\Sigma_T, \tag{5.3}$$

$$g(x+2L,y,t)=g(x,y,t),\qquad (x,y,t)\in Q_T, \tag{5.4}$$

$$g(x,y,0)=g_0(x,y),\qquad (x,y)\in\mathbb{R}\times(0,e), \tag{5.5}$$

$$f_t-\alpha\gamma f_{xx}-g(x,0,t)=F_4(x,t)\quad \text{in } \Sigma_T, \tag{5.6}$$

$$f(x+2L,t)=f(x,t),\qquad (x,t)\in\Sigma_T, \tag{5.7}$$

$$f(x,0)=f_0(x),\qquad x\in\mathbb{R}, \tag{5.8}$$

where T is a fixed (arbitrary) positive number, $F_1\in C_{\text{per}}^{\lambda,\lambda/2}(\overline{Q}_T)$, $F_i\in C_{\text{per}}^{1+\lambda,(1+\lambda)/2}(\overline{\Sigma}_T)$, $i=2,\dots,4$, for some $\lambda\in(0,1)$, $f_0\in C_{\text{per}}^{3+\lambda}(\mathbb{R})$, $g_0\in C_{\text{per}}^{2+\lambda}(\mathbb{R}\times[0,e])$ and satisfies the compatibility conditions

$$\begin{aligned} g_{oy}(x,e) &=F_3(x,0), \qquad x\in\mathbb{R}\\ (g_0-g_{0y})(x,0)+\alpha\gamma f_0''(x) &=F_2(x,0), \qquad x\in\mathbb{R}.\end{aligned}$$

THEOREM 5.1. *Problem* (5.1)–(5.8) *has a unique solution*

$$(g,f)\in C_{\text{per}}^{2+\lambda,1+\lambda/2}(\overline{Q}_T)\times C_{\text{per}}^{3+\lambda,(3+\lambda)/2}(\overline{\Sigma}_T).$$

Furthermore, there exists a constant $C=C(T)$ *such that*

$$\begin{aligned}\|g\|_{Q_T}^{(2+\lambda)}+\|f\|_{\Sigma_T}^{(3+\lambda)} \le C\Big(&\|g_0\|_{\mathbb{R}\times(0,e)}^{(2+\lambda)}+\|f_0\|_{\mathbb{R}}^{(3+\lambda)}\\ &+\|F_1\|_{Q_T}^{(\lambda)}+\|F_2\|_{\Sigma_T}^{(1+\lambda)}+\|F_3\|_{\Sigma_T}^{(1+\lambda)}+\|F_4\|_{\Sigma_T}^{(1+\lambda)}\Big).\end{aligned}$$

Proof. We consider the mapping $\mathcal{L}:f\to\mathcal{L}(f)=h$ defined as follows. Given $f\in C_{\text{per}}^{3+\lambda,(3+\lambda)/2}(\overline{\Sigma}_T)$ with $f(0)=f_0$ (where we use the notation $f(0)=f|_{t=0}$) we consider (5.1)–(5.5). By [9, Thm. 5.3 p. 320] this problem has a unique solution

$g \in C^{2+\lambda,1+\lambda/2}(\overline{Q}_T)$, and the uniqueness property together with the periodicity of the data implies that $g \in C^{2+\lambda,1+\lambda/2}_{\text{per}}(\overline{Q}_T)$. We then compute $h = \mathcal{L}f \in C^{3+\lambda,(3+\lambda)/2}_{\text{per}}(\overline{\Sigma}_T)$ as the unique solution of (5.6)–(5.8) [9, Thm. 5.1 p. 320]. The following results on continuous dependence then hold:

$$\|g\|^{(2+\lambda)}_{Q_T} \leq C_1 \Big(\|g_0\|^{(2+\lambda)}_{\mathbb{R}\times(0,e)} + \|F_1\|^{(\lambda)}_{Q_T} + \|F_2\|^{(1+\lambda)}_{\Sigma_T} + \|F_3\|^{(1+\lambda)}_{\Sigma_T} + \|f\|^{(3+\lambda)}_{\Sigma_T} \Big), \tag{5.9}$$

$$\|h\|^{(3+\lambda)}_{\Sigma_T} \leq C_2 \Big(\|f_0\|^{(3+\lambda)}_{\mathbb{R}} + \|F_4\|^{(1+\lambda)}_{\Sigma_T} + \|g\|^{(2+\lambda)}_{Q_T} \Big), \tag{5.10}$$

where C_1 and C_2 only depend on T, λ and on the parameters e, α, and γ.

Clearly if f is a fixed point of $\mathcal{L}$, i.e., if $\mathcal{L}(f) = f$ and if g is the solution of the problem (5.1)–(5.5), then (f, g) satisfies (5.1)–(5.8). Moreover, $\mathcal{L}$ is an affine mapping from the affine subspace $\{f \in C^{3+\lambda,(3+\lambda)/2}_{\text{per}}(\overline{\Sigma}_T), f(0) = f_0\}$ of $C^{3+\lambda,(3+\lambda)/2}_{\text{per}}(\overline{\Sigma}_T)$ into itself. The linear mapping $\mathcal{M}$ corresponding to $\mathcal{L}$, which is such that

$$\mathcal{L}(f) - \mathcal{L}(\bar{f}) = \mathcal{M}(f - \bar{f}), \tag{5.11}$$

is defined on

$$\overset{\circ}{C}{}^{3+\lambda,(3+\lambda)/2}_{\text{per}}(\overline{\Sigma}_T) := \{f \in C^{3+\lambda,(3+\lambda)/2}_{\text{per}}(\overline{\Sigma}_T), f(0) = 0\}$$

by

$$f \to \mathcal{M}(f) = k,$$

where k is the solution of

$$\begin{cases} k_t - \alpha\gamma k_{xx} = w(x,0,t) & \text{in } \Sigma_T, \\ k(x+2L,t) = k(x,t) & \text{in } \Sigma_T, \\ k(x,0) = 0 & x \in \mathbb{R}, \end{cases} \tag{5.12}$$

where w is the solution of

$$\begin{cases} w_t - \Delta w = 0 & \text{in } Q_T, \\ w - w_y = -\alpha\gamma f_{xx} & \text{on } \Sigma_T, \\ w_y(x,e,t) = 0, & (x,t) \in \Sigma_T, \\ w(x+2L,y,t) = w(x,y,t) & \text{in } Q_T, \\ w(x,y,0) = 0, & (x,y) \in \mathbb{R} \times (0,e). \end{cases} \tag{5.13}$$

In what follows we shall prove that $(Id - \mathcal{M})$ is a continuous invertible mapping from $\overset{\circ}{C}{}^{3+\lambda,(3+\lambda)/2}_{\text{per}}(\overline{\Sigma}_T)$ into itself; next we show that this property is equivalent to the existence of a unique fixed point of $\mathcal{L}$. Adding $f - \bar{f}$ to both sides of equality (5.11) yields

$$\mathcal{L}(f) - \mathcal{L}(\bar{f}) + f - \bar{f} - \mathcal{M}(f - \bar{f}) = f - \bar{f}$$

so that

$$(Id - \mathcal{M})(f - \bar{f}) = \mathcal{L}(\bar{f}) - \bar{f} - \mathcal{L}(f) + f.$$

Thus solving

$$\mathcal{L}(f) = f$$

is equivalent to solving

$$(Id - \mathcal{M})(f - \bar{f}) = \mathcal{L}(\bar{f}) - \bar{f},$$

which gives

$$f = \bar{f} + (Id - \mathcal{M})^{-1}(\mathcal{L}(\bar{f}) - \bar{f}). \tag{5.14}$$

In a first step we shall prove that $\mathcal{M}$ is a compact operator and in a second step that Ker $(Id - \mathcal{M}) = \{0\}$. It will then follow from the Fredholm alternative (see, for instance, Brezis [1, Thm. 6.1]) that the operator $(Id - \mathcal{M})$ is invertible.

(i) $\mathcal{M}$ is a compact operator from $\overset{\circ}{C}{}_{\text{per}}^{3+\lambda,(3+\lambda)/2}(\overline{\Sigma}_T)$ into itself.

Let f be given in $\overset{\circ}{C}{}_{\text{per}}^{3+\lambda,(3+\lambda)/2}(\overline{\Sigma}_T)$; by [9] the solution w of (5.13) belongs to $\overset{\circ}{C}{}_{\text{per}}^{2+\lambda,(2+\lambda)/2}(\overline{Q}_T)$ so that also by [9] the solution k of (5.12) belongs to $\overset{\circ}{C}{}_{\text{per}}^{4+\lambda,2+\lambda/2}(\overline{\Sigma}_T)$. Thus $\mathcal{M}$ is a bounded operator from $\overset{\circ}{C}{}_{\text{per}}^{3+\lambda,(3+\lambda)/2}(\overline{\Sigma}_T)$ into $\overset{\circ}{C}{}_{\text{per}}^{4+\lambda,2+\lambda/2}(\overline{\Sigma}_T)$. Using, furthermore, the compactness of the embedding from

$$\overset{\circ}{C}{}_{\text{per}}^{4+\lambda,2+\lambda/2}(\overline{\Sigma}_T) \quad \text{into } \overset{\circ}{C}{}_{\text{per}}^{3+\lambda,(3+\lambda)/2}(\overline{\Sigma}_T)$$

we conclude that $\mathcal{M}$ is a compact operator from $\overset{\circ}{C}{}_{\text{per}}^{3+\lambda,(3+\lambda)/2}(\overline{\Sigma}_T)$ into itself.

(ii) Ker $(Id - \mathcal{M}) = \{0\}$.

Let $f \in \overset{\circ}{C}{}_{\text{per}}^{3+\lambda,(3+\lambda)/2}(\overline{\Sigma}_T)$ be such that $(Id - \mathcal{M})f = 0$, i.e., $\mathcal{M}f = f$. Then if w is the unique solution of (5.13), f satisfies the system

$$\begin{cases} f_t - \alpha\gamma f_{xx} = w(x, 0, t) & \text{in } \Sigma_T, \\ f(x + 2L, t) = f(x, t) & \text{in } \Sigma_T, \\ f(x, 0) = 0, & x \in \mathbb{R}. \end{cases} \tag{5.15}$$

Next we multiply the differential equation in (5.15) first by f and then by $-\alpha\gamma f_{xx}$ and integrate by parts on $(-L, L)$. This gives

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\int_{-L}^{L} f^2(x,t)dx + \alpha\gamma \int_{-L}^{L} f_x^2(x,t)dx \\ &= \int_{-L}^{L} w(x,0,t)f(x,t)dx \end{aligned} \tag{5.16}$$

and

$$\begin{aligned} &\frac{\alpha\gamma}{2}\frac{d}{dt}\int_{-L}^{L} f_x^2(x,t)dx + \alpha^2\gamma^2 \int_{-L}^{L} f_{xx}^2(x,t)dx \\ &= -\alpha\gamma \int_{-L}^{L} w(x,0,t)f_{xx}(x,t)dx. \end{aligned} \tag{5.17}$$

Now we multiply the differential equation in (5.13) by w and deduce from the boundary condition that w satisfies

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_0^e\int_{-L}^L w^2(x,y,t)\,dx\,dy+\int_0^e\int_{-L}^L(\operatorname{grad}w(x,y,t))^2dx\,dy\\ =\int_{-L}^L(w(x,0,t)+\alpha\gamma f_{xx}(x,t))w(x,0,t)dx.\end{aligned}\tag{5.18}$$

Using Young's inequality we deduce from (5.16) that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_{-L}^L f^2(x,t)dx+\alpha\gamma\int_{-L}^L f_x^2(x,t)dx\\ \le\int_{-L}^L w^2(x,0,t)dx+\frac{1}{4}\int_{-L}^L f^2(x,t)dx.\end{aligned}\tag{5.19}$$

Adding (5.17), (5.18), and (5.19) yields

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_{-L}^L f^2dx+\alpha\gamma\int_{-L}^L f_x^2dx+\frac{\alpha\gamma}{2}\frac{d}{dt}\int_{-L}^L f_x^2dx+\alpha^2\gamma^2\int_{-L}^L f_{xx}^2dx\\ +\frac{1}{2}\frac{d}{dt}\int_0^e\int_{-L}^L w^2dx\,dy+\int_0^e\int_{-L}^L(\operatorname{grad}w)^2dx\,dy\\ \le 2\int_{-L}^L w^2(x,0,t)dx+\frac{1}{4}\int_{-L}^L f^2dx.\end{aligned}\tag{5.20}$$

Finally we use the following result which we shall prove later.

LEMMA 5.2. *Let Ω be a bounded Lipschitzian domain in $\mathbb{R}^2$; for all $\varepsilon>0$ there exists a positive constant C_ε such that*

$$\int_{\partial\Omega}w^2(s)ds\le\varepsilon\iint_\Omega(\operatorname{grad}w)^2dx+C_\varepsilon\iint_\Omega w^2dx$$

for all $w\in H^1(\Omega)$.

We set $\varepsilon=\frac{1}{4}$ to deduce that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\left(\int_{-L}^L f^2dx+\alpha\gamma\int_{-L}^L f_x^2dx+\int_0^e\int_{-L}^L w^2dx\,dy\right)\\ \le\frac{1}{4}\int_{-L}^L f^2dx+2C_{1/4}\int_0^e\int_{-L}^L w^2dx\,dy.\end{aligned}\tag{5.21}$$

The inequality above is of the form $dY/dt \leq CY$, with

$$Y(t) = \frac{1}{2}\int_{-L}^{L} f^2 dx + \frac{\alpha\gamma}{2}\int_{-L}^{L} f_x^2 dx + \frac{1}{2}\int_{0}^{e}\int_{-L}^{L} w^2 dx\, dy,$$

and since $Y(0) = 0$ we conclude that $Y(t) = 0$ for all t. This implies that f and w are identically equal to zero and, therefore, that $Id - \mathcal{M}$ is an invertible mapping.

Finally the continuous dependence of the fixed point f of $\mathcal{L}$ on the data $F_i, i = 1, 2, 3, 4, f_0$, and g_0 can be seen as follows. Choose $\bar{f}(x,t) = f_0(x)$ in (5.14); then f is given by

$$f = f_0 + (Id - \mathcal{M})^{-1}(\mathcal{L}(f_0) - f_0), \tag{5.22}$$

and we deduce from (5.9), (5.10) that

$$\begin{aligned} \|\mathcal{L}(f_0)\|_{\Sigma_T}^{(3+\lambda)} \leq & C_3 \Big(\|f_0\|_{\mathbb{R}}^{(3+\lambda)} + \|F_4\|_{\Sigma_T}^{(1+\lambda)} \\ & + \|g_0\|_{\mathbb{R}\times(0,e)}^{(2+\lambda)} + \|F_1\|_{Q_T}^{(\lambda)} + \|F_2\|_{\Sigma_T}^{(1+\lambda)} + \|F_3\|_{\Sigma_T}^{(1+\lambda)} \Big), \end{aligned} \tag{5.23}$$

where we have used the fact that $\|f_0\|_{\Sigma_T}^{(3+\lambda)} = \|f_0\|_{\mathbb{R}}^{(3+\lambda)}$. Therefore, since $(Id - \mathcal{M})^{-1}$ is a bounded operator, we have that

$$\|f\|_{\Sigma_T}^{(3+\lambda)} \leq C_4 \left(\|f_0\|_{\mathbb{R}}^{(3+\lambda)} + \|\mathcal{L}(f_0)\|_{\Sigma_T}^{(3+\lambda)} \right). \tag{5.24}$$

The result of Theorem 5.1 then follows from (5.9), (5.23), and (5.24).

Proof of Lemma 5.2. We have that

$$\int_{\partial\Omega} w^2(s) ds = \|w\|_{H^0(\partial\Omega)}^2 \leq C\|w\|_{H^{1/2+\delta}(\Omega)}^2$$

for all $\delta > 0$. Next we use the following result given, for instance, by [10, p. 47]:

$$H^{1/2+\delta}(\Omega) = [H^1(\Omega), H^0(\Omega)]_{1/2-\delta}$$

so that

$$\|w\|_{H^{1/2+\delta}(\Omega)} \leq C(\delta)\|w\|_{H^1(\Omega)}^{1/2+\delta}\|w\|_{L^2(\Omega)}^{1/2-\delta}.$$

Choose for instance $\delta = \frac{1}{4}$. Then

$$\|w\|_{H^{3/4}(\Omega)} \leq C\|w\|_{H^1(\Omega)}^{3/4}\|w\|_{L^2(\Omega)}^{1/4},$$

that is,

$$\|w\|_{H^{3/4}(\Omega)}^2 \leq C(\|w\|_{H^1(\Omega)}^2)^{3/4}(\|w\|_{L^2(\Omega)}^2)^{1/4}.$$

Using the inequality

$$a^{3/4}b^{1/4} = (\varepsilon a)^{3/4}\left(\frac{b}{\varepsilon^3}\right)^{1/4} \leq \varepsilon a + \frac{b}{\varepsilon^3}$$

for all $a, b, \varepsilon > 0$, we deduce that

$$\|w\|_{H^{3/4}(\Omega)}^2 \leq C(\tilde{\varepsilon}\|w\|_{H^1(\Omega)}^2 + \|w\|_{L^2(\Omega)}^2/\tilde{\varepsilon}^3),$$

which implies the result of Lemma 5.2.

6. Uniqueness of the solution. We can prove a uniqueness result that holds in a larger class than that for which the existence result of §7 can be proven.

THEOREM 6.1. *There exists at most one solution (g, f) of Problem P such that $g \in L^\infty(0,T; W^{2,\infty}_{\text{per}}(\mathbb{R} \times (0,e)))$, $g_t \in L^\infty(Q_T)$ and $f \in L^\infty(0,T; W^{3,\infty}_{\text{per}}(\mathbb{R}))$, $f_t \in L^\infty(\Sigma_T)$.*

Proof. A detailed proof of this result would make the paper too bulky and we shall only sketch it briefly. The main idea is, similar to the linear problem, to obtain a differential inequality for the difference of two solutions. In fact, we can prove that the function

$$Y(t) = \int_0^e \int_{-L}^{L} (g_1 - g_2)^2 + \int_{-L}^{L} (f_1 - f_2)^2_x + \int_{-L}^{L} (f_1 - f_2)^2$$

satisfies an inequality of the form

$$\frac{dY}{dt} \leq MY \quad \text{with } Y(0) = 0,$$

where M can be chosen (using numerous integrations by parts and interpolation inequalities) so as to depend only on the bounds of g_1, g_2, f_1, f_2 in the spaces given in the statement of Theorem 6.1. This, of course, is enough to prove Theorem 6.1.

7. Existence. Upon omitting the tildas, we can rewrite the nonlinear problem P in the form

$$g_t - \Delta g = f_x^2 g_{yy} - 2f_x g_{xy} - f_{xx} g_y + f_t g_y \quad \text{in } Q_T, \tag{7.1}$$

$$\begin{aligned}(g - g_y)(x,0,t) + \alpha\gamma f_{xx}(x,t) &= \alpha(e^{-\gamma f_{xx}/(1+f_x^2)^{3/2}} - 1 + \gamma f_{xx}) \\ &+ ((1+f_x^2)^{1/2} - 1)g_y(x,0,t) - \frac{f_x}{(1+f_x^2)^{1/2}} g_x(x,0,t), \quad (x,t) \in \Sigma_T,\end{aligned} \tag{7.2}$$

$$g_y(x,e,t) = (1 - (1+f_x^2)^{1/2}) g_y(x,e,t) + \frac{f_x}{(1+f_x^2)^{1/2}} g_x(x,e,t), \ (x,t) \in \Sigma_T, \tag{7.3}$$

$$g(x+2L,y,t) = g(x,y,t), \qquad (x,y,t) \in Q_T, \tag{7.4}$$

$$g(x,y,0) = g_0(x,y), \qquad (x,y) \in \mathbb{R} \times (0,e), \tag{7.5}$$

$$\begin{aligned}f_t - \alpha\gamma f_{xx} - g(x,0,t) &= -\alpha(1+f_x^2)^{1/2}\left(e^{-\gamma f_{xx}/(1+f_x^2)^{3/2}} - 1 + \gamma\frac{f_{xx}}{(1+f_x^2)^{1/2}}\right) \\ &+ ((1+f_x^2)^{1/2} - 1)g(x,0,t) \quad \text{in } \Sigma_T,\end{aligned} \tag{7.6}$$

$$f(x+2L,t) = f(x,t), \qquad (x,t) \in \Sigma_T, \tag{7.7}$$

$$f(x,0) = f_0(x), \qquad x \in \mathbb{R}. \tag{7.8}$$

Remark that the left-hand sides of the equations in the linear problem (5.1)–(5.8) and in problem (7.1)–(7.8) coincide. Here again we suppose that

$$g_0 \in C^{2+\lambda}_{\text{per}}(\mathbb{R} \times [0, e]), f_0 \in C^{3+\lambda}_{\text{per}}(\mathbb{R})$$

and that g_0 satisfies compatibility conditions, namely, (7.2) and (7.3) in which f and g are replaced by f_0 and g_0.

Next we define

$$F_1(f,g) = f_x^2 g_{yy} - 2 f_x g_{xy} - f_{xx} g_y + f_t g_y, \tag{7.9}$$

$$\begin{aligned} F_2(f,g) =& \alpha(e^{-\gamma f_{xx}/(1+f_x^2)^{3/2}} - 1 + \gamma f_{xx}), \\ &+ ((1+f_x^2)^{1/2} - 1) g_y(x,0,t) - \frac{f_x}{(1+f_x^2)^{1/2}} g_x(x,0,t), \end{aligned} \tag{7.10}$$

$$F_3(f,g) = (1 - (1+f_x^2)^{1/2}) g_y(x,e,t) + \frac{f_x}{(1+f_x^2)^{1/2}} g_x(x,e,t), \tag{7.11}$$

$$\begin{aligned} F_4(f,g) =& -\alpha(1+f_x^2)^{1/2}\left(e^{-\gamma f_{xx}/(1+f_x^2)^{3/2}} - 1 + \gamma \frac{f_{xx}}{(1+f_x^2)^{1/2}}\right) \\ &+ ((1+f_x^2)^{1/2} - 1) g(x,0,t), \end{aligned} \tag{7.12}$$

and prove the following preliminary lemma.

LEMMA 7.1. *Let $\delta_0 > 0$ be arbitrary.*

(i) *There exists a positive constant $K_1 = K_1(\delta_0)$ such that for each $\delta \in (0, \delta_0)$*

$$\|F_1(f,g)\|^{\lambda}_{Q_T} + \sum_{i=2}^{4} \|F_i(f,g)\|^{(1+\lambda)}_{\Sigma_T} \le K_1(\|f\|^{(3+\lambda)}_{\Sigma_T} + \|g\|^{(2+\lambda)}_{Q_T})^2$$

for all $f \in C^{3+\lambda,(3+\lambda)/2}(\overline{\Sigma}_T)$ and $g \in C^{2+\lambda,1+\lambda/2}(\overline{Q}_T)$ satisfying $\|f\|^{(3+\lambda)}_{\Sigma_T} + \|g\|^{(2+\lambda)}_{Q_T} \le \delta$.

(ii) *There exists a constant $K_2 = K_2(\delta_0)$ such that for each $\delta \in (0, \delta_0)$ and for all $f, h \in C^{3+\lambda,(3+\lambda)/2}(\overline{\Sigma}_T)$, $g, k \in C^{2+\lambda,1+\lambda/2}(\overline{Q}_T)$ satisfying*

$$\max(\|f\|^{(3+\lambda)}_{\Sigma_T} + \|g\|^{(2+\lambda)}_{Q_T}, \|h\|^{(3+\lambda)}_{\Sigma_T} + \|k\|^{(2+\lambda)}_{Q_T}) \le \delta,$$

there holds

$$\begin{aligned} &\|F_1(f,g) - F_1(h,k)\|^{(\lambda)}_{Q_T} + \sum_{i=2}^{4} \|F_i(f,g) - F_i(h,k)\|^{(1+\lambda)}_{\Sigma_T} \\ &\le K_2 \delta \left(\|f - h\|^{(3+\lambda)}_{\Sigma_T} + \|g - k\|^{(2+\lambda)}_{Q_T}\right). \end{aligned}$$

Proof. (i) For F_i the property directly follows from Lemma 4.1 (i). Next we show the property for F_2. By Lemma 4.1 (i), (ii)

$$\left\| \frac{f_{xx}}{(1+f_x^2)^{j/2}} \right\|^{(1+\lambda)}_{\Sigma_T} \le C \|f_{xx}\|^{(1+\lambda)}_{\Sigma_T} \left(\|(1+f_x^2)^{j/2}\|^{1+\lambda}_{\Sigma_T}\right)^2 \quad \text{for } j = 1, 3.$$

Using also Lemma 4.1 (iii) we deduce that if $f \in C^{3+\lambda,(3+\lambda)/2}(\overline{\Sigma}_T)$ there exists a constant $K_1 = K_1(\delta_0)$ such that

$$\left\| \frac{f_{xx}}{(1+f_x^2)^{j/2}} \right\|_{\Sigma_T}^{(1+\lambda)} \leq K_1 \|f\|_{\Sigma_T}^{(3+\lambda)} \quad \text{for } j = 1, 3.$$

Similarly we can check that there exists $K_2 = K_2(\delta_0)$ such that

$$\|(1+f_x^2)^{1/2} - 1\|_{\Sigma_T}^{(1+\lambda)} \leq K_2 \|f\|_{\Sigma_T}^{(3+\lambda)}$$

and $K_3 = K_3(\delta_0)$ such that

$$\left\| \frac{1}{(1+f_x^2)^{1/2}} - \frac{1}{(1+f_x^2)^{3/2}} \right\|_{\Sigma_T}^{(1+\lambda)} \leq K_3 \left(\|f\|_{\Sigma_T}^{(3+\lambda)} \right)^2.$$

Thus using also (4.1) and Lemma 4.2 we deduce that

$$\begin{aligned} \|F_2(f,g)\|_{\Sigma_T}^{(1+\lambda)} &\leq K_4 \left\{ (\|f\|_{\Sigma_T}^{(3+\lambda)})^2 + \|f\|_{\Sigma_T}^{(3+\lambda)} \|g\|_{Q_T}^{(2+\lambda)} \right\} \\ &\leq K_5 \left(\|f\|_{\Sigma_T}^{(3+\lambda)} + \|g\|_{Q_T}^{(2+\lambda)} \right)^2, \end{aligned}$$

and that similar results hold for F_3 and F_4.

(ii) The property is obvious for F_1. For instance,

$$\begin{aligned} &\|f_x^2 g_{yy} - h_x^2 k_{yy}\|_{Q_T}^{(\lambda)} \\ &\quad \leq \|(f_x^2 - h_x^2) g_{yy}\|_{Q_T}^{(\lambda)} + \|h_x^2 (g_{yy} - k_{yy})\|_{Q_T}^{(\lambda)} \\ &\quad \leq C \left(\|g_{yy}\|_{Q_T}^{(\lambda)} \|f_x + h_x\|_{\Sigma_T}^{(\lambda)} \|f_x - h_x\|_{\Sigma_T}^{(\lambda)} + (\|h_x\|_{\Sigma_T}^{(\lambda)})^2 \|g_{yy} - k_{yy}\|_{Q_T}^{(\lambda)} \right) \\ &\quad \leq C\delta^2 \left(\|f - h\|_{\Sigma_T}^{(3+\lambda)} + \|g - k\|_{Q_T}^{(2+\lambda)} \right). \end{aligned}$$

As for F_2 we first consider the term

$$\begin{aligned} F &= \left\| \frac{f_x}{(1+f_x^2)^{1/2}} g_x(x,0,t) - \frac{h_x}{(1+h_x^2)^{1/2}} k_x(x,0,t) \right\|_{\Sigma_T}^{(1+\lambda)} \\ &\leq C \left\| \frac{f_x}{(1+f_x^2)^{1/2}} \right\|_{\Sigma_T}^{(1+\lambda)} \|g_x(x,0,t) - k_x(x,0,t)\|_{\Sigma_T}^{(1+\lambda)} \\ &\quad + C \|k_x(x,0,t)\|_{\Sigma_T}^{(1+\lambda)} \left\| \frac{f_x}{(1+f_x^2)^{1/2}} - \frac{h_x}{(1+h_x^2)^{1/2}} \right\|_{\Sigma_T}^{(1+\lambda)}. \end{aligned}$$

Since

$$\frac{f_x}{(1+f_x^2)^{1/2}} - \frac{h_x}{(1+h_x^2)^{1/2}} = \left(\int_0^1 \frac{d\theta}{(1 + (\theta f_x + (1-\theta) h_x)^2)^{3/2}} \right) (f_x - h_x),$$

we deduce that

$$\begin{aligned}
&\left\| \frac{f_x}{(1+f_x^2)^{1/2}} - \frac{h_x}{(1+h_x^2)^{1/2}} \right\|_{\Sigma_T}^{(1+\lambda)} \\
&\quad \le C_1 \left(\int_0^1 d\theta \|(1+(\theta f_x + (1-\theta)h_x)^2)^{3/2}\|_{\Sigma_T}^{(1+\lambda)} \right)^2 \|f_x - h_x\|_{\Sigma_T}^{(1+\lambda)} \\
&\quad \le C_2 \left(\|(1+f_x^2+h_x^2)^{3/2}\|_{\Sigma_T}^{(1+\lambda)} \right)^2 \|(f-h)_x\|_{\Sigma_T}^{(1+\lambda)} \\
&\quad \le C_3 \left(\|(1+f_x^2)^{3/2}\|_{\Sigma_T}^{(1+\lambda)} + \|(1+h_x^2)^{3/2}\|_{\Sigma_T}^{(1+\lambda)} \right)^2 \|(f-h)_x\|_{\Sigma_T}^{(1+\lambda)} \\
&\quad \le C \|(f-h)_x\|_{\Sigma_T}^{(1+\lambda)}
\end{aligned}$$

so that

$$\begin{aligned}
F &\le K \left(\|f_x\|_{\Sigma_T}^{(1+\lambda)} \|(g-k)_x(x,0,t)\|_{\Sigma_T}^{(1+\lambda)} \right. \\
&\qquad \left. + \|k_x(x,0,t)\|_{\Sigma_T}^{(1+\lambda)} \|(f-h)_x\|_{\Sigma_T}^{(1+\lambda)} \right) \\
&\le K\delta \left(\|f-h\|_{\Sigma_T}^{(3+\lambda)} + \|g-k\|_{Q_T}^{(2+\lambda)} \right).
\end{aligned}$$

The term $((1+f_x^2)^{1/2}-1)g_y(x,0,t) - ((1+h_x^2)^{1/2}-1)k_y(x,0,t)$ can be estimated in a similar way. Finally let us consider the term

$$E = \chi\left(\frac{-f_{xx}}{(1+f_x^2)^{3/2}} \right) + \frac{f_x^2 f_{xx}}{(1+f_x^2)^{3/2}}$$

in (4.1). Then we can estimate the difference

$$E_2 = \frac{f_x^2 f_{xx}}{(1+f_x^2)^{3/2}} - \frac{h_x^2 h_{xx}}{(1+h_x^2)^{3/2}}$$

as it has been done above with F. Finally we consider

$$E_1 = \chi\left(\frac{-f_{xx}}{(1+f_x^2)^{3/2}} \right) + \chi\left(\frac{h_{xx}}{(1+h_x^2)^{3/2}} \right).$$

We have that

$$\left\| \frac{f_{xx}}{(1+f_x^2)^{3/2}} \right\|_{\Sigma_T}^{(1+\lambda)} \le C \|f_{xx}\|_{\Sigma_T}^{(1+\lambda)} \left(\|(1+f_x^2)^{3/2}\|_{\Sigma_T}^{(1+\lambda)} \right)^2 \le K\delta$$

so that by Lemma 4.3,

$$\begin{aligned}
E_1 &\le K'\delta \left\| \frac{f_{xx}}{(1+f_x^2)^{3/2}} - \frac{h_{xx}}{(1+h_x^2)^{3/2}} \right\|_{\Sigma_T}^{(1+\lambda)} \\
&\le K'\delta \|f-h\|_{\Sigma_T}^{(3+\lambda)}.
\end{aligned}$$

A similar proof can be given for F_3 and F_4.

Next we state the main result of this section.

THEOREM 7.2. *There exists a positive constant $\rho_0 = \rho_0(T)$ such that if the initial data satisfy the condition*

$$\|f_0\|_{\mathbb{R}}^{(3+\lambda)} + \|g_0\|_{\mathbb{R}\times(0,e)}^{(2+\lambda)} \leq \rho_0, \tag{7.13}$$

then there exists a constant $\rho > \rho_0$ such that (7.1)–(7.8) *has a unique solution (f, g) in $C_{\text{per}}^{3+\lambda,(3+\lambda)/2}(\overline{\Sigma}_T) \times C_{\text{per}}^{2+\lambda,1+\lambda/2}(\overline{Q}_T)$ satisfying $\|f\|_{\Sigma_T}^{(3+\lambda)} + \|g\|_{Q_T}^{(2+\lambda)} \leq \rho$.*

Proof. We consider the mapping $\mathcal{T}$

$$(f, g) \to \mathcal{T}(f, g) = (f', g')$$

defined as follows: (f', g') is the solution of the linear problem (5.1)–(5.8) with initial conditions f_0 and g_0 and where the right-hand sides $F_i, i = 1, \ldots, 4$ are given by the formulas (7.9)–(7.12).

By Lemma 7.1 (i) and Theorem 5.1 the mapping $\mathcal{T}$ is well defined from the affine space

$$A = \left\{(f, g) \in C_{\text{per}}^{3+\lambda,(3+\lambda)/2}(\overline{\Sigma}_T) \times C_{\text{per}}^{2+\lambda,1+\lambda/2}(\overline{Q}_T), f(0) = f_0, g(0) = g_0\right\}$$

into itself. Next we define the norm $\|\cdot\|$ on A by

$$\|(f, g)\| = \|f\|_{\Sigma_T}^{(3+\lambda)} + \|g\|_{Q_T}^{(2+\lambda)}.$$

In what follows we prove that if the constant ρ_0 in (7.13) is small enough, then the mapping $\mathcal{T}$ has a unique fixed point in a certain ball. It follows from the study of the linear case (Theorem 5.1) that

$$\begin{aligned}\|\mathcal{T}(f, g)\| &= \|f'\|_{\Sigma_T}^{(3+\lambda)} + \|g'\|_{Q_T}^{(2+\lambda)} \\ &\leq \mathcal{C}\left(\|f_0\|_{\mathbb{R}}^{(3+\lambda)} + \|g_0\|_{\mathbb{R}\times(0,e)}^{(2+\lambda)} + \|F_1\|_{Q_T}^{(\lambda)} + \sum_{i=2}^{4}\|F_i\|_{\Sigma_T}^{(1+\lambda)}\right),\end{aligned} \tag{7.14}$$

where $\mathcal{C}$ is a constant that depends on T. This implies that

$$\begin{aligned}\|\mathcal{T}(f, g) - \mathcal{T}(h, k)\| \leq \mathcal{C}\Bigg(&\|F_1(f, g) - F_1(h, k)\|_{Q_T}^{(\lambda)} \\ &+ \sum_{i=2}^{4}\|F_i(f, g) - F_i(h, k)\|_{\Sigma_T}^{(1+\lambda)}\Bigg).\end{aligned} \tag{7.15}$$

Also using Lemma 7.1 (ii), we deduce that if (f, g) and (h, k) satisfy

$$\|(f, g)\| \leq \delta, \|(h, k)\| \leq \delta,$$

then

$$\|\mathcal{T}(f, g) - \mathcal{T}(h, k)\| \leq \mathcal{C}K_2\delta\|(f - h, g - k)\|;$$

so that if

$$\delta < \frac{1}{\mathcal{C}K_2} \tag{7.16}$$

and if $\mathcal{B}(0,\delta)$ denotes the ball of center zero and radius δ, $\mathcal{T}$ is a strict contraction from the closed convex set $\mathcal{B}(0,\delta)\cap A$ into A.

Finally we deduce from (7.14) and from Lemma 7.1 (i) that if

$$\mathcal{C}\left(\|f_0\|_{\mathbb{R}}^{(3+\lambda)} + \|g_0\|_{\mathbb{R}\times(0,e)}^{(2+\lambda)} + K_1\left(\|f\|_{\Sigma_T}^{(3+\lambda)} + \|g\|_{Q_T}^{(2+\lambda)}\right)^2\right) \le \delta, \tag{7.17}$$

then $\mathcal{T}$ maps the set $\mathcal{B}(0,\delta)\cap A$ into itself. In turn, (7.17) is implied by the condition

$$\widetilde{\mathcal{C}}(\rho_0 + K_1\delta^2) \le \delta \tag{7.18}$$

with

$$\widetilde{\mathcal{C}} = \max(\mathcal{C}, 1).$$

Note that (7.18) implies that $\delta > \rho_0$ so that the set $\mathcal{B}(0,\delta)\cap A$ is nonempty. Choose, for example,

$$\delta = \frac{1}{2\widetilde{\mathcal{C}}K} \quad \text{and} \quad \rho_0 = \frac{1}{4\widetilde{\mathcal{C}}^2K}$$

with $K = \max(K_1, K_2)$. Then we can check that (7.16) and (7.18) are satisfied.

Acknowledgment. The authors are very indebted to M. Cournil, G. Santarini, and C. Bataillon for suggesting this problem, showing them the physical derivation, and for many fruitful discussions. D. Hilhorst wishes to thank Professor H. F. Weinberger for very inspiring comments.

REFERENCES

[1] H. Brezis, *Analyse Fonctionnelle*, Théorie et Applications, Masson, Paris 1987.

[2] Xinfu Chen, *Generation and propagation of interfaces in reaction-diffusion systems*, IMA Preprint Series # 708, September 1990.

[3] X. Y. Chen, *Dynamics of interfaces in reaction-diffusion systems*, Hiroshima Math. J., 21 (1991), pp. 47–83.

[4] F. Conrad, D. Hilhorst, and T. I. Seidman, *Well-posedness of a moving boundary problem arising in a dissolution-growth process*, Nonlinear Analysis TMA, 15 (1990), pp. 445–465.

[5] M. Cournil, private communication.

[6] J. Duchon and R. Robert, *Evolution d'une interface par capillarité et diffusion de volume* I. *Existence locale en temps*, Ann. Inst. H. Poincaré, 1 (1984), pp. 361–378.

[7] D. Hilhorst, F. Issard-Roch, and T. I. Seidman, *On a reaction-diffusion equation with a free boundary: the case of an unbounded domain*, in Proceedings of Free Boundary Problems, Theory and Applications, Montreal, Canada, 1990, to appear.

[8] F. Kaleydjian and M. Cournil, *Stability of steady states in some solid-liquid systems*, React. Solids, 2 (1986), pp. 1–21.

[9] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Translations of Mathematical Monographs, 23, American Mathematical Society, Providence, RI, 1968.

[10] R. Temam, *Navier–Stokes Equations*, North-Holland, Amsterdam, 1987.

SIAM J. MATH. ANAL.
Vol. 24, No. 2, pp. 317-326, March 1993

GLOBAL BEHAVIOR OF POSITIVE SOLUTIONS TO NONLINEAR DIFFUSION PROBLEMS WITH NONLINEAR ABSORPTION THROUGH THE BOUNDARY*

NOEMÍ WOLANSKI†

Abstract. The author studies the global behavior of positive solutions to $u_t = \Delta\phi(u)$ in $B_R \times (0, T)$; $\partial\phi(u)/\partial\nu = f(u)$ on $S_R \times (0, T)$; and $u(x, 0) = u_0(x)$ in B_R. Here f and ϕ are increasing functions of u, positive for u positive which go to infinity as $u \to \infty$, and $0 < \phi'(u) < \infty$ in R. (ϕ may degenerate in either way at $u = 0$ if $u_0 \geqq \delta \geqq 0$.) It is proved that when $\phi'(u) \geqq C > 0$ in R there exists a unique global solution when f is sublinear and finite time blowup occurs when $\int^\infty (ds/f(s)) < \infty$. On the other hand, if one allows $\phi'(u)$ to go to zero as $u \to \infty$ finite time blowup may occur with f being sublinear. The author gives precise relations between ϕ and f that guarantee global existence or finite time blowup. Results are illustrated with a couple of examples.

Key words. blowup, nonlinear diffusion, nonlinear absorption

AMS(MOS) subject classifications. 35K60, 35K55, 35B35

1. Introduction.

In this paper we study the global behavior of positive solutions to a nonlinear diffusion problem with a nonlinear absorption-like boundary condition.

The problem we analyze is

$$
(1.1)\qquad
\begin{aligned}
&u_t = \Delta\phi(u) && \text{in } B_R \times (0, T),\\
&\frac{\partial\phi(u)}{\partial\nu} = f(u) && \text{on } S_R \times (0, T),\\
&u(x, 0) = u_0(x) && \text{in } B_R,
\end{aligned}
$$

where $B_R = \{|x| < R\}$; $S_R = \{|x| = R\}$ and u_0 is a smooth nonnegative function such that

$$
(1.2)\qquad \frac{\partial\phi(u_0)}{\partial\nu} = f(u_0) \quad \text{on } S_R.
$$

ϕ and f are increasing functions of u that are positive for u positive together with their derivatives, and which go to infinity as u goes to infinity.

The case $\phi(u) = u$ is well understood. In fact, classical results show that when $f(u)/u$ is bounded there is a global solution for any $u_0 \in L^\infty(B_R)$. On the other hand, we have proved in [LMW] that every solution blows up in finite time in L^∞-norm if $\int^\infty (ds/f(s)) < \infty$.

The general case is not well understood. For a proof of local existence and uniqueness of weak solutions when u_0 is allowed to vanish and $\phi'(0) = 0$, we refer to [A]. In fact, in this case we do not expect to have classical solutions since moving fronts could exist on which ∇u is discontinuous.

In this work we are interested in the local boundedness or finite time blowup of the solutions and not in their behavior near $u = 0$. This is the reason why we will assume throughout the paper that $0 < \phi'(u) < \infty$ in $\mathbf{R}$. (If this is not the case we will assume that $u \geqq \delta > 0$, which is guaranteed by u_0 being larger than δ and modifying ϕ on the interval $(-\delta/2, \delta/2)$ so as to have $0 < \phi' < \infty$ everywhere.) In this case the

* Received by the editors October 2, 1991; accepted for publication (in revised form) July 9, 1992.

† Departmento de Matemática, Facultad de Ciencias Exactas, Universidad de Buenos Aires, (1428) Buenos Aires, Argentina. Member of CONICET. This author's work was partially supported by the Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota and Fundación Antorchas, Argentina.

compatibility condition (1.2) implies that any radial solution is a classical one if $f \in C^1(\mathbf{R})$ (see Proposition 2.1).

We find that the results valid in the case $\phi(u)=u$ continue to hold as long as $\phi'(u) \geqq C > 0$. Instead, when $\liminf_{(u\to\infty)} \phi'(u)=0$, blowup may occur with f being sublinear.

In fact, if ϕ is concave, $\liminf_{u\to\infty} f(u)\sqrt{\phi'(u)}/\phi(u) > 0$, and $\int^\infty (\sqrt{\phi'(s)}/f(s))\,ds < \infty$ every solution blows up in finite time. On the other hand, if $(f(u)\sqrt{\phi'(u)}/\phi(u))$ is bounded every solution exists globally.

We want to point out that as in [LMW] these results hold for every initial datum u_0 (not necessarily radial). Blowup results for some u_0 by concavity methods in the one-dimensional case can be found in [A].

Let us illustrate our results with some examples.

(a) Let $f(u)=u^q$, $\phi(u)=u^m$. There are two cases ($u_0 \geqq \delta > 0$).

(i) $m \geqq 1$.

If $q > 1$ every solution blows up in finite time.

If $q \leqq 1$ every solution exists globally.

(ii) $0 < m < 1$.

If $q > (m+1)/2$ every solution blows up in finite time.

If $q \leqq (m+1)/2$ every solution exists globally.

These results were first found in the one-dimensional case by Filo in [F]. Although the case $0 < m < 1$, $q=(m+1)/2$ was not covered by his proofs.

(b) Let $\phi(u)=\log(u+1)$.

If $f(u) \leqq C\sqrt{u+1}\log(u+1)$ every solution exists globally.

If $f(u) \geqq C\sqrt{u+1}(\log(u+1))^{1+\varepsilon}$ as $u\to\infty$ for some $\varepsilon>0$, every solution blows up in finite time.

Our results generalize [F] in two ways. First, we allow for general nonlinearities. Second, our domain is an N-dimensional ball (and even any simply connected two-dimensional smooth domain; see Remark 3.1).

In fact, some of the proofs in [F] require very special initial data, whose existence is proved only in the one-dimensional case with homogeneous ϕ and f.

For other papers dealing with the balance between reaction and diffusion or convection, or two reactions of opposite sign, we refer to [F], [ChLS], [LPSSt], [L], [LS], [LSm], [LMW2], and [ChFQ].

The rest of the paper is organized as follows. In § 2 we state and prove the global existence and in § 3 the finite time blowup results.

2. Global existence results. We state without proof the following result on local existence and continuation of classical solutions.

PROPOSITION 2.1. *Let $\phi \in C^{2+\alpha}$, $f \in C^1$ and $u_0 \in C^{2+\alpha}(\bar{B}_R)$, $u_0=u_0(r)$ satisfy the compatibility condition* (1.2). *If $\phi'(u)>0$ in $\mathbf{R}$ there exists $T>0$ and $u \in C^{2+\alpha,1+(\alpha/2)}(\bar{B}_R \times [0,T))$ solution of* (1.1). *If u cannot be continued as a classical solution beyond time T we have*

$$\limsup_{t\to T} |u(R,t)| = +\infty.$$

The proof is a trivial modification of the one in [LMW] for the heat equation. It is based on a fixed point argument in the space $C^{1+\alpha,(1+\alpha)/2}(\bar{B}_R \times [0,T])$ and the existence of classical solutions to the third boundary value problem in the linear case (see [LMW, Thm. 1.2, Cor. 1.2]).

For a proof of local existence, continuation, uniqueness, and comparison of weak solutions in the nonsymmetric case we refer to [A].

It is clear from the continuation results that in order to prove global existence it is enough to show that any local solution is a priori bounded on any finite time interval.

In order to state this result we need to consider two different situations, $\phi'(u) \geqq C > 0$ or $0 < \phi'(u) \leqq C$.

THEOREM 2.1. *Let Ω be a smooth domain. Let u be a weak solution of*

$$\begin{aligned} &u_t = \Delta\phi(u) && \text{in } \Omega \times (0, T), \\ &\frac{\partial\phi(u)}{\partial\nu} = f(u) && \text{on } \partial\Omega \times (0, T), \\ &u(x, 0) = u_0(x) && \text{in } \Omega. \end{aligned}$$

Let us assume that $\phi'(u) \geqq C > 0$ and $f(u)/(1+u) \leqq M < \infty$. There exists a constant L depending only on T and u_0 such that

$$u(x, t) \geqq L \quad \text{in } \Omega \times (0, T).$$

Proof. We follow here ideas in [LSU], [S], and [A]. Let $k > \|u_0\|_\infty$. Then,

$$(u-k)_t = \operatorname{div}\,(\phi'(u)\nabla(u-k)).$$

By taking $(u-k)^+ = \max\,(u-k, 0)$ as a test function, we obtain

$$\begin{aligned} \frac{1}{2}\int_\Omega ((u-k)^+)^2(x,t)\,dx &= \frac{1}{2}\int_\Omega ((u_0-k)^+)^2(x)\,dx \\ &\quad + \int_0^t \int_\Omega \operatorname{div}\,(\phi'(u)\nabla(u-k))(u-k)^+(x,s)\,dx\,ds \\ &= -\int_0^t \int_\Omega \phi'(u)|\nabla(u-k)^+|^2(x,s)\,dx\,ds \\ &\quad + \int_0^t \int_{\partial\Omega} f(u)(u-k)^+\,d\sigma\,ds \\ &\leqq -C\int_0^t \int_\Omega |\nabla(u-k)^+|^2\,dx\,ds + M\int_0^t \int_{\partial\Omega} u(u-k)^+\,d\sigma\,ds \\ &\quad + M\int_0^t \int_{\partial\Omega} (u-k)^+\,d\sigma\,ds. \end{aligned}$$

Thus, if $k \geqq 1$,

$$\begin{aligned} \frac{1}{2}\int_\Omega ((u-k)^+)^2(x,t)\,dx &+ C\int_0^t \int_\Omega |\nabla(u-k)^+|^2\,dx\,ds \\ &\leqq M\int_0^t \int_{\partial\Omega} ((u-k)^+)^2\,d\sigma\,ds + 2Mk\int_0^t \int_{\partial\Omega} (u-k)^+\,d\sigma\,ds. \end{aligned}$$

On the other hand, if $\mu(k) = \text{meas}\,\{(x,s) \in \Omega \times (0,T) \colon u(x,s) > k\}$,

$$\begin{aligned} \int_0^t \int_{\partial\Omega} (u-k)^+\,d\sigma\,ds &\leqq K\left\{\int_0^t \int_\Omega (u-k)^+\,dx\,ds + \int_0^t \int_\Omega |\nabla(u-k)^+|\,dx\,ds\right\} \\ &\leqq \frac{1}{2Mk}\int_0^t \int_\Omega ((u-k)^+)^2\,dx\,ds \\ &\quad + \frac{C}{2Mk}\int_0^t \int_\Omega |\nabla(u-k)^+|^2\,dx\,ds + \frac{M}{2}K^2(1+C^{-1})k\mu(k). \end{aligned}$$

Since

$$\int_{\partial\Omega}((u-k)^+)^2\,d\sigma \leqq K\left\{\int_\Omega ((u-k)^+)^2\,dx + \int_\Omega |\nabla((u-k)^+)^2|\,dx\right\}$$
$$= K\left\{\int_\Omega ((u-k)^+)^2\,dx + 2\int_\Omega (u-k)^+|\nabla(u-k)^+|\,dx\right\},$$

for any $\varepsilon > 0$,

$$\int_0^t\int_{\partial\Omega}((u-k)^+)^2\,d\sigma\,ds \leqq \varepsilon\int_0^t\int_\Omega |\nabla(u-k)^+|^2\,dx\,ds + C_\varepsilon\int_0^t\int_\Omega ((u-k)^+)^2\,dx\,ds.$$

So, we have

$$\frac{1}{2}\int_\Omega ((u-k)^+)^2(x,t)\,dx + \frac{C}{4}\int_0^t\int_\Omega |\nabla(u-k)^+|^2\,dx\,ds$$
$$\leqq K_1k^2\mu(k) + K_2\int_0^t\int_\Omega ((u-k)^+)^2\,dx\,ds.$$

Using Gronwall's inequality we see that the right-hand side is bounded for $t \leqq T$ by $K_3k^2\mu(k)$.

On the other hand, for any function $v(x, t)$,

$$\|v\|^2_{L^{2((N+2)/N)}(\Omega\times(0,t))} \leqq K_4\left\{\sup_{0\leqq s\leqq t}\|v(\cdot, s)\|^2_{L^2(\Omega)} + \|\nabla_v\|^2_{L^2(\Omega\times(0,t))}\right\};$$

thus,

$$\left(\int_0^T\int_\Omega ((u-k)^+)^{2((N+2)/2)}\,dx\,dt\right)^{N/(N+2)} \leqq K_5k^2\mu(k).$$

Let $h > k \geqq (\|u_0\|_{L^\infty} + 1)$. The latter inequality implies

$$(h-k)^2\mu(h)^{N/(N+2)} \leqq K_5k^2\mu(k), \tag{2.1}$$

and a simple modification of Stampacchia's argument ([KS], page 63), which is necessary due to the presence of k^2 on the right-hand side of (2.1), shows that there exists a constant $L > 0$ such that $\mu(L) = 0$.

In fact, let $k = \|u_0\|_{L^\infty} + 1$ in (2.1). Since for any k, $\mu(k) \leqq T \operatorname{meas}\Omega$ we have for $h > \|u_0\|_{L^\infty} + 1$,

$$\mu(h)^{N/(N+2)} \leqq \frac{K_5(\|u_0\|_{L^\infty}+1)^2 T \operatorname{meas}\Omega}{(h - \|u_0\|_{L^\infty} - 1)^2}.$$

Let $\beta = (N+2)/N$ and $\sigma = 2\beta/(\beta-1)$. From the inequality above we see that we can take $k_0 = k_0(T, N, \Omega, \|u_0\|_{L^\infty})$ large enough so as to have $k_0 > (1 + \|u_0\|_{L^\infty})$ and

$$\sqrt{K_5}\,2^{2+\sigma/2}\mu(k_0)^{(\beta-1)/2\beta} \leqq 1.$$

Set $d = K_5^\beta 2^{\beta(4+\sigma)}k_0^{2\beta}\mu(k_0)^{\beta-1}$ and $k_n = k_0 + d^{1/2\beta} - (d^{1/2\beta}/2^n)$. We claim that

$$\mu(k_n) \leqq \frac{\mu(k_0)}{2^{n\sigma}}. \tag{2.2}$$

In fact, for $h > k$,

$$\mu(h) \leqq \frac{K_5^\beta k^{2\beta}\mu(k)^\beta}{(h-k)^{2\beta}}.$$

Thus

$$\mu(k_{n+1}) \leqq \frac{K_5^\beta k_n^{2\beta} \mu(k_n)^\beta 2^{2\beta(n+1)}}{d}.$$

For $n=0$ this gives

$$\mu(k_1) \leqq \frac{K_5^\beta k_0^{2\beta} \mu(k_0)^{\beta-1} 2^{2\beta}}{K_5^\beta 2^{\beta(4+\sigma)} k_0^{2\beta} \mu(k_0)^{\beta-1}} \mu(k_0) = \frac{\mu(k_0)}{2^{\beta(2+\sigma)}} \leqq \frac{\mu(k_0)}{2^\sigma}.$$

Assuming that (2.2) is true,

$$\begin{aligned} \mu(k_{n+1}) &\leqq \frac{K_5^\beta k_n^{2\beta} 2^{2\beta(n+1)}}{d} \frac{\mu(k_0)^\beta}{2^{n\sigma\beta}} \\ &= \left(\frac{k_n}{k_0}\right)^{2\beta} 2^{2\beta(n+1)-\beta(4+\sigma)-n\sigma\beta} \mu(k_0) \\ &\leqq 2^{2n\beta-\beta\sigma-n\beta\sigma} \mu(k_0), \end{aligned}$$

since by construction $k_n/k_0 \leqq 2$. By the choice of σ,

$$2n\beta - \sigma\beta(n+1) \leqq -\sigma(n+1).$$

Taking limit as $(n \to \infty)$ in (2.2) we get

$$\mu(k_0 + d^{1/2\beta}) = 0.$$

This means that $u(x, t) \leqq k_0 + d^{1/2\beta} \leqq 2k_0$ in $\Omega \times (0, T)$.

Let us now analyze the case $0 < \phi'(u) \leqq C$ in $\mathbf{R}$.

THEOREM 2.2. *Let f and ϕ be strictly increasing smooth functions such that $0 < \phi'(u) \leqq C$. Assume ϕ concave or $f(u)/\phi(u)$ nondecreasing. Let $\sqrt{\phi'(u)}(f(u)/\phi(u)) \leqq M$ for $u \in \mathbf{R}$. For any positive classical solution of* (1.1) *there exists a constant L depending on u_0 and T such that*

$$u(x, t) \leqq L \quad \textit{in } B_R \times (0, T).$$

Proof. We prove it first for radially symmetric, radially increasing solutions such that $u_t \geqq 0$. The result follows by comparison with these solutions.

Assume first that $f(u)/\phi(u)$ is nondecreasing. Let $\beta(v) = \phi^{-1}(v)$ and $\theta = C^{-1}$, then $\beta'(v) \geqq \theta$ for $v \in \mathbf{R}$. v is a solution of

$$\text{(2.3)} \qquad \begin{cases} \beta(v)_t = \Delta v & \text{in } B_R \times (0, T) \\ v_r(R, t) = g(v(R, t)) & \text{on } (0, T), \\ v(x, 0) = v_0(x), \end{cases}$$

where $v_0 = \beta(u_0)$, $g(v) = f(\beta(v))$. From the hypothesis we have that

$$\text{(2.4)} \qquad \left(\frac{g(v)}{v}\right)^2 \leqq M\beta'(v).$$

Let

$$h(r, t) = 1 + \exp\left(-\frac{1}{TR} \int_0^t \int_0^r \int_0^s \frac{g(v(\eta, \tau))}{v(\eta, \tau)} d\eta \, ds \, d\tau\right);$$

then for $0 \leqq r \leqq R$, $0 < t \leqq T$,

$$\text{(2.5)} \qquad 2 \geqq h \geqq 1,$$

$$0 \geqq h_r \geqq -\frac{g(v(r,t))}{v(r,t)} h, \tag{2.6}$$

$$\frac{h_r^2}{h} \leqq \left(\frac{g(v(r,t))}{v(r,t)}\right)^2 h, \tag{2.7}$$

$$\Delta h \geqq -\frac{n}{R}\frac{g(v(r,t))}{v(r,t)} h, \tag{2.8}$$

$$h_t \leqq 0. \tag{2.9}$$

These inequalities are trivially verified by direct differentiation using the fact that $g(v(r,t))/v(r,t)$ is nondecreasing in r and t and the estimate

$$\frac{h_r}{r} \geqq -\frac{1}{R}\frac{g(v)}{v} h.$$

Let $V(r,t) = e^{-Kt}v(r,t)h(r,t)$. Then

$$\beta'(v)V_t - \Delta V + 2\frac{h_r}{h}V_r + \left\{\left(K - \frac{h_t}{h}\right)\beta'(v) + \frac{\Delta h}{h} - 2\frac{h_r^2}{h^2}\right\}V = 0 \quad \text{in } B_R \times (0,T),$$

$$V_r(R,t) - \left\{\frac{h_r}{h}(R,t) + \frac{g(v(R,t))}{v(R,t)}\right\}V(R,t) = 0, \qquad t > 0,$$

$$V(r,0) = v_0(r)h(r,0).$$

From (2.4) and the fact that $\beta'(v) \geqq \theta > 0$ we find that for another constant $\bar{M}$,

$$\frac{g(v)}{v} \leqq \bar{M}\beta'(v). \tag{2.10}$$

By (2.4)-(2.10) we see that there exists $K > 0$ such that

$$Kh(r,t)\beta'(v(r,t)) \geqq h_t\beta'(v) - \Delta h + 2\frac{h_r^2}{h}.$$

Also by (2.5),

$$\frac{h_r}{h}(R,t) + \frac{g(v(R,t))}{v(R,t)} \geqq 0.$$

By the maximum principle we see that

$$V(r,t) \leqq \|V_0\|_{L^\infty} \leqq 2\|v_0\|_{L^\infty}.$$

Thus,

$$v(r,t) \leqq 2\|v_0\|_{L^\infty} e^{Kt}.$$

Now, if $g(v)/v$ is not nondecreasing but β is convex, let $G(v)$ be defined by

$$\frac{G(v)}{v} = \max_{0 \leqq s \leqq v} \frac{g(s)}{s},$$

then

$$\left(\frac{G(v)}{v}\right)^2 \leqq M\beta'(v).$$

Let w be the solution of

$$\begin{aligned} &\beta(w)_t = \Delta w && \text{in } B_R \times (0, T), \\ &w_r(R, t) = G(w(R, t)), && t > 0, \\ &w(x, 0) = v_0(x) && \text{in } B_R. \end{aligned}$$

Then, since $G(v) \geqq g(v)$ in $\mathbf{R}$, $w \geqq v$ in $B_R \times (0, T)$. Thus v is bounded.

The general case with no further assumptions on u follows by comparison with the solution of (1.1) that starts with initial value $u_1(|x|)$ satisfying

$$\begin{gathered} u_1 \geqq \|u_0\|_{L^\infty}, \\ u_{1r} \geqq 0, \\ \Delta\phi(u_1) \geqq a > 0, \\ \phi(u_1)_r(R) = f(u_1(R)), \end{gathered}$$

in B_R. The solution that starts with initial value u_1 will have nonnegative time and radial derivative. We state this result as Lemma 2.1. In Lemma 2.2 we show that we can always find such a function.

LEMMA 2.1. *Let u_1 be a radially symmetric classical solution of* (1.1). *If $u_{1r}(x, 0) \geqq 0$ and $\Delta\phi(u_1)(x, 0) \geqq a > 0$, u_1 satisfies*

$$\left.\begin{aligned} u_{1r} \geqq 0 \\ u_{1t} \geqq 0 \end{aligned}\right. \quad \textit{in } B_R \times (0, T).$$

Proof. Let $0 < \tau < T$; then u_1, u_{1r} and u_{1t} are bounded in $B_R \times (0, \tau)$. Thus, there exist constants θ, $K > 0$ such that $\phi'(u_1)$, $|\phi''(u_1)| \leqq \theta$ in $B_R \times (0, \tau)$, and if $\beta = \phi^{-1}$, $v = \phi(u_1)$,

$$K \geqq -\frac{\beta''(v)}{\beta'(v)} v_t.$$

Letting $w = e^{-Kt} v_r(r, t)$ it is easy to see that w satisfies

$$\beta'(v) w_t - \Delta w = -\left[\frac{n-1}{r^2} + \beta''(v) v_t + K\beta'(v)\right] w,$$

so it is clear that w cannot attain a negative interior minimum. On the other hand, $w \geqq 0$ on $t = 0$ and $r = R$. Thus $w \geqq 0$ everywhere.

In order to prove that u_{1t} is nonnegative we need a different argument. In fact the $C^{2+\alpha, 1+(\alpha/2)}$ regularity of u_1 does not allow us to use flux boundary conditions for u_t. Following the approach in [LMW] we consider the functions $u_\varepsilon(x, t) = u_1(x, t + \varepsilon)$. Since $u_{1t}(x, 0) \geqq a > 0$ and u_{1t} is continuous there exists $\varepsilon_0 > 0$ such that $u_\varepsilon(x, 0) \geqq u_1(x, 0)$ in B_R for $0 < \varepsilon < \varepsilon_0$. It is easy to see that u_ε is a solution of (1.1). Also since u_1 is bounded in $B_R \times (0, \tau + \varepsilon_0)$ if $\tau + \varepsilon_0 < T$ we may assume that ϕ and ϕ' are bounded from above and below. Classical comparison principles then give us the inequality

$$u(x, t + \varepsilon) \geqq u(x, t)$$

in $B_R \times (0, \tau)$ for every $\varepsilon < \varepsilon_0$. This is $u_t \geqq 0$.

LEMMA 2.2. *Let γ, $L > 0$. There exist radially symmetric functions u_1, $u_2 \in C^{2+\alpha}(\bar{B}_R)$ such that $u_i \geqq 0$, $u_{ir} \geqq 0$, $\phi(u_i)_{rr} \geqq a$, $\phi(u_i)_r(R) = f(u_i(R))$, and $u_1 \geqq L$, $u_2 \leqq \gamma$ in B_R.*

Proof. Let $w_i \geqq 0$ smooth such that $w_i(0)=0$, $w_i'(r) \geqq a$, $w_i(R)=f(c_i)$, and

$$\int_0^R w_i(r)\, dr \leqq \tfrac{1}{2}\phi(c_i),$$

where

(i) c_1 is large enough as to have $\phi(c_1) \geqq 2\phi(L)$;

(ii) $c_2=\gamma$.

Then

$$u_i(x)=\phi^{-1}\left(\phi(c_i)-\int_{|x|}^R w_i(r)\, dr\right)$$

are the functions we are looking for.

3. Blowup results. In this section we prove the blowup results stated in § 1. As in the previous section we have to consider two different situations, depending on whether or not $\phi'(u) \geqq C > 0$.

THEOREM 3.1. *Let f and ϕ be strictly increasing smooth functions with $\int^\infty ds/f(s) < \infty$. Every positive solution of* (1.1) *blows up in L^∞-norm in finite time.*

Proof. The proof follows step by step that of Proposition 3.1 in [LMW], so we omit it here. The blowup result is proven first by very simple arguments for radially symmetric, radially increasing classical solutions and obtained afterwards for any positive solution by comparison by using Lemma 2.2 and the fact that any nonnegative solution is strictly positive in $\bar{B}_R$ for $t>0$.

A different argument has to be used to get a sharper blowup result when $\liminf_{u\to\infty} \phi'(u)=0$.

THEOREM 3.2. *Let f and ϕ be strictly increasing smooth functions such that ϕ is concave,* $\liminf_{u\to\infty} f(u)\sqrt{\phi'(u)}/\phi(u)>0$, *and* $\int^\infty (\sqrt{\phi'(s)}/f(s))\, ds < \infty$. *Every positive solution of* (1.1) *blows up in L^∞-norm in finite time.*

Proof. As in the proof of Theorem 3.1 we first consider radially symmetric, radially increasing classical solutions and get the general result by comparison with these solutions by using the construction in Lemma 2.2.

Let us first find some energy identities to be used in the proof. First multiply (1.1) by $\phi(u)_t$ and integrate to get

$$\int_0^t\int_{B_R} \phi'(u)u_t^2 = -\frac{1}{2}\int_0^t\int_{B_R} \frac{\partial}{\partial t}(|\nabla\phi(u)|^2)+\int_0^t\int_{S_R} \phi'(u)f(u)u_t\, d\sigma\, dt.$$

Let $\Phi(u)=\int_0^u \sqrt{\phi'(s)}\, ds$; $F(u)=\int_0^u \phi'(s)f(s)\, ds$. Then,

$$\int_0^t\int_{B_R} |\Phi(u)_t|^2+\frac{1}{2}\int_{B_R} |\nabla\phi(u)|^2=\frac{1}{2}\int_{B_R} |\nabla\phi(u_0)|^2+\int_{S_R} F(u)\, d\sigma-\int_{S_R} F(u_0)\, d\sigma. \tag{3.1}$$

Next multiply (1.1) by $\phi(u)$ and integrate to get

$$\int_0^t\int_{B_R} \phi(u)u_t=-\int_0^t\int_{B_R} |\nabla\phi(u)|^2+\int_0^t\int_{B_R} \phi(u)f(u)\, d\sigma\, dt.$$

Let $\Theta(u)=\int_0^u \phi(s)\, ds$. Then,

$$\int_{B_R} \Theta(u)(x,t)\, dx+\int_0^t\int_{B_R} |\nabla\phi(u)|^2=\int_{B_R} \Theta(u_0)\, dx+\int_0^t\int_{S_R} \phi(u)f(u)\, d\sigma\, dt. \tag{3.2}$$

Let us now compute

$$\begin{aligned}\Phi(u)_t\phi(u)_r &= \sqrt{\phi'(u)}u_t\phi(u)_r\\ &= \sqrt{\phi'(u)}\phi(u)_r\left(\phi(u)_{rr}+\frac{n-1}{r}\phi(u)_r\right)\\ &\geqq \sqrt{\phi'(u)}\phi(u)_r\phi(u)_{rr}\\ &= \frac{1}{2}\sqrt{\phi'(u)}\frac{\partial}{\partial r}|\phi(u)_r|^2.\end{aligned}$$

Thus

$$\int_{B_R}\Phi(u)_t\phi(u)_r \geqq -\frac{1}{4}\int_{B_R}\frac{\phi''(u)}{\sqrt{\phi'(u)}}u_r|\phi(u)_r|^2 + \frac{1}{2}R^{n-1}\omega_{n-1}\sqrt{\phi'(u(R,t))}(f(u(R,t)))^2.$$

Now, since $\phi''\leqq 0$ and $u_r\geqq 0$,

$$\int_{B_R}\Phi(u)_t\phi(u)_r \geqq \frac{1}{2}R^{n-1}\omega_{n-1}\sqrt{\phi'(u(R,t))}(f(u(R,t)))^2.$$

Applying Schwarz's inequality to the left-hand side and integrating in time we get

$$\int_0^t\sqrt{\phi'(u(R,s))}(f(u(R,s)))^2\,ds \leqq \varepsilon\int_0^t\int_{B_R}|\phi(u)_r|^2 + K\int_0^t\int_{B_R}|\Phi(u)_t|^2.$$

By (3.2) the right-hand side is bounded by

$$C_1 + K\int_0^t\int_{B_R}|\Phi(u)_t|^2 + \varepsilon\omega_{n-1}R^{n-1}\int_0^t\phi(u(R,s))f(u(R,s))\,ds,$$

where C_1 is a constant depending on u_0.

By hypothesis, $\phi(u)f(u)\leqq M + C\sqrt{\phi'(u)}(f(u))^2$; so we have

$$\int_0^t\sqrt{\phi'(u(R,s))}(f(u(R,s)))^2\,ds \leqq C_2 + \bar{K}\int_0^t\int_{B_R}|\Phi(u)_t|^2.$$

By (3.1) the right-hand side is bounded by

$$C_3 + KF(u(R,t)).$$

This is

$$\int_0^t\sqrt{\phi'(u(R,s))}(f(u(R,s)))^2\,ds \leqq C_3 + \hat{K}F(u(R,t)). \tag{3.3}$$

Let $G(u)=\sqrt{\phi'(u)}(f(u))^2$, $g(t)=\int_0^t G(u(R,s))\,ds$, and $H=G\circ F^{-1}$. By (3.3), since H is increasing,

$$g'(t) = G(u(R,t)) = H(F(u(R,t))) \geqq H\left(\frac{g(t)-C_3}{\hat{K}}\right). \tag{3.4}$$

Assume now that $\int^\infty(du/H(u))<\infty$. We deduce that g blows up in finite time. Therefore, u cannot exist globally. This implies that u must blow up in finite time.

The condition $\int^\infty(du/H(u))<\infty$ translates into $\int^\infty(\sqrt{\phi'(u)}/f(u))\,du<\infty$. Since this is our hypothesis the proof is finished.

REMARK 3.1. In the two-dimensional case our results extend to any simply connected smooth domain Ω. In fact, under a conformal mapping of Ω onto the disc B_1, a solution $v(y, t)$ of

$$\begin{aligned} &v_t = \Delta\phi(v) && \text{in } \Omega \times (0, T), \\ &\frac{\partial}{\partial \nu}\phi(v) = f(v) && \text{on } \partial\Omega \times (0, T), \\ &v(y, 0) = v_0(y) && \text{in } \Omega \end{aligned} \tag{3.5}$$

is transformed into a solution $u(x, t)$ of

$$\begin{aligned} &u_t = \alpha(x)\Delta\phi(u) && \text{in } B_1 \times (0, T), \\ &\alpha(x)\phi(u)_r = f(u) && \text{on } S_1 \times (0, T), \\ &u(x, 0) = u_0(x) && \text{in } B_1, \end{aligned} \tag{3.6}$$

where $0 < \alpha \leqq \alpha(x) \leqq \beta < \infty$ in $\bar{B}$. Since solutions of (3.6) can be compared to time increasing solutions of

$$\begin{aligned} &u_t = \alpha\Delta\phi(u) && \text{in } B_1 \times (0, T), \\ &\beta\phi(u)_r = f(u) && \text{on } S_1 \times (0, T), \\ &u(x, 0) = u_0(x) && \text{in } B_1, \end{aligned} \tag{3.7}$$

and all our results are valid for (3.7), we can extend our theorems to problem (3.5) for any smooth simply connected two-dimensional Ω (see [LMW] for the details).

Acknowledgments. This work was finished while the author was visiting the Institute for Mathematics and Its Applications, University of Minnesota. I wish to thank this institution for its hospitality.

REFERENCES

[A] J. ANDERSON, *Local existence and uniqueness of solutions of degenerate parabolic equations*, Comm. Partial Differential Equations, 16 (1991), pp. 105–143.

[ChFQ] M. CHIPOT, F. FILA, AND P. QUITTNER, *Stationary solutions, blow up and convergence to stationary solutions for semilinear parabolic equations with nonlinear boundary conditions*, Acta Math. Univ. Comenian, 60 (1991), pp. 35–103.

[ChLS] T. F. CHEN, H. LEVINE, AND P. SACKS, *Analysis of a convective reaction-diffusion equation*, Nonlinear Anal. T.M.A., 12 (1988), pp. 1349–1370.

[F] J. FILO, *Diffusivity versus absorption through the boundary*, J. Differential Equations, to appear.

[KS] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variation Inequalities and Their Applications*, Academic Press, New York, 1980.

[L] H. LEVINE, *Stability and instability of solutions to Burger's equation with a semilinear boundary condition*, SIAM J. Math. Anal., 19 (1988), pp. 312–336.

[LMW] J. LÓPEZ GÓMEZ, V. MÁRQUEZ, AND N. WOLANSKI, *Blow up results and localization of blow up points for the heat equation with a nonlinear boundary condition*, J. Differential Equations, 92 (1991), pp. 384–401.

[LMW2] ———, *Global behavior of positive solutions to a semilinear equation with a nonlinear flux condition*, IMA preprint #830, Univ. of Minnesota, Minneapolis, MN, submitted.

[LPSSt] H. LEVINE, L. PAYNE, P. SACKS, AND B. STRAUGHAN, *Analysis of a convective reaction-diffusion equation* (II), SIAM J. Math. Anal., 20 (1989), pp. 133–147.

[LS] H. LEVINE AND P. SACKS, *Some existence and nonexistence theorems for solutions of degenerate parabolic equations*, J. Differential Equations, 52 (1984), pp. 135–161.

[LSm] H. LEVINE AND R. SMITH, *A potential well theory for the heat equation with a nonlinear boundary condition*, Math. Methods Appl. Sci., 9 (1987), pp. 127–136.

[LSU] O. LADYZHENSKAYA, V. SOLONNIKOV, AND C. URALCEVA, *Linear and Quasilinear equations of Parabolic type*, Transl. Math. Monographs American Mathematical Society, 23 (1968), Providence, RI.

[S] P. SACKS, *Existence and regularity of solutions of inhomogeneous porous medium type equations*, TSR #2214, Math. Research Center, Madison, WI.

SIAM J. MATH. ANAL.
Vol. 24, No. 2, pp. 327–340, March 1993

005

BUCKLING EIGENVALUES FOR A CLAMPED PLATE EMBEDDED IN AN ELASTIC MEDIUM AND RELATED QUESTIONS*

BERNHARD KAWOHL†, HOWARD A. LEVINE‡, AND WALDEMAR VELTE§

Abstract. This paper considers the dependence of the sum of the first m eigenvalues of three classical problems from linear elasticity on a physical parameter in the equation. The paper also considers eigenvalues $\gamma_i(a)$ of a clamped plate under compression, depending on a lateral loading parameter a; $\Lambda_i(a)$, the Dirichlet eigenvalues of the elliptic system describing linear elasticity depending on a combination a of the Lamé constants, and eigenvalues $\Gamma_i(a)$ of a clamped vibrating plate under tension, depending on the ratio a of tension and flexural rigidity. In all three cases $a \in [0, \infty)$. The analysis of these eigenvalues and their dependence on a gives rise to some general considerations on singularly perturbed variational problems.

Key words. eigenvalue, asymptotic, parameter dependence, plate equation, elasticity, singular perturbation

AMS(MOS) subject classifications. 35J50, 35J55, 35P15, 49G05, 49G20

Introduction. Let, for $i = 1, 2, \ldots, \gamma_i$ be the eigenvalues of the equation for the clamped plate under compression, Γ_i be the eigenvalues for the equations of linear elasticity, and Λ_i be the eigenvalues for the equation for the vibrating clamped plate under tension. Briefly, our first result says that $\sum_{i=1}^{m} \gamma_i(a)$ and $\sum_{i=1}^{m} \Gamma_i(a)$ are strictly concave functions of a, while $\sum_{i=1}^{m} \Lambda_i(a)$ is concave. Moreover and in particular

$$\begin{aligned}
&\lim_{a\to\infty} \gamma_1(a) = +\infty \quad \text{and} \quad \lim_{a\to\infty} \frac{\gamma_1(a)}{\sqrt{a}} = 2,\\
&\lim_{a\to\infty} \Lambda_1(a) < \infty,\\
&\lim_{a\to\infty} \Gamma_1(a) = +\infty \quad \text{but} \quad \lim_{a\to\infty} \frac{\Gamma_1(a)}{a} = \lambda_1.
\end{aligned}$$

(Here λ_1 is the first Dirichlet eigenvalue for the Laplacian which is also known as the first eigenvalue for the fixed membrane.) The graphs of these functions are sketched in Figs. 1, 2, and 4 along with the previously known upper and lower bounds. The plan of the paper is as follows. In §1 we discuss $\gamma_i(a)$, $\Lambda_i(a)$ and $\Gamma_i(a)$. We use some ideas of [10], [11], [12] to obtain some of our results. In §§2 and 3 we consider generalizations, first to abstract linear problems and then to nonlinear problems. Throughout the paper $\{\lambda_i\}_{i\in N}$ denotes the ordered sequence of eigenvalues of the problem

$$\begin{aligned}
\Delta\psi + \lambda\psi &= 0 \quad \text{in } \Omega,\\
\psi &= 0 \quad \text{on } \partial\Omega,
\end{aligned}$$

while $\{\psi_j\}_{j\in N}$ denotes the corresponding sequence of orthonormal eigenfunctions.

*Received by the editors July 15, 1991; accepted for publication (in revised form) June 9, 1992.

†SBF 123, Universität Heidelberg, Im Neuenheimer Feld 294, D 6900 Heidelberg, Germany. This author was supported in part by the Deutsche Forschungsgemeinschaft (DFG) via SBF 123 and a Heisenberg award.

‡Mathematics Department, Iowa State University, Ames, Iowa 50011. This author was supported by the Deutsche Forschungsgemeinschaft (DFG) via SBF 123 as well as National Science Foundation grant DMS-8822788.

§Institut für Angewandte Mathematik, Universität Würzburg, D8700 Würzburg, Germany.

1. The first eigenvalue of a clamped plate under compression. Let $\Omega \subset I\!R^N$ be a domain with smooth boundary and let $a \geq 0$ be a parameter. Consider the eigenvalue problem

$$\tag{1} \begin{aligned} \Delta\Delta u + au + \gamma(a)\Delta u &= 0 \quad \text{in } \Omega, \\ u = \frac{\partial u}{\partial n} &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where $a > 0$ is given and represents the elasticity constant of a medium surrounding the plate. The function u stands for the transverse displacement [9] and $\gamma_1(a)$ is the minimal compression at which the plate exhibits buckling. Payne established the following inequality in [9]

$$\tag{2} \max\{\gamma_1(0),\ 2\sqrt{a}\} \leq \ \gamma_1(a) \ \leq \gamma_1(0) + \frac{a}{\lambda_1}.$$

(Indeed, he showed that (2) holds for each eigenvalue $\gamma_i(a)$ of (1).) Moreover, Levine and Protter derived the lower bound

$$\tag{3} \sum_{i=1}^{m} \gamma_i(a) \geq \frac{4\pi^2 N m^{(1+2/N)}}{(N+2)(\omega_N V)^{2/N}}$$

in [6]. Here ω_N denotes the surface area of the unit ball in $I\!R^N$ and V denotes the volume of Ω.

THEOREM 1. *The function* $F(a) = \sum_{i=1}^{m} \gamma_i(a)$ *is strictly concave and strictly increasing in* a *on* $[0, \infty)$.

For the proof we use the variational characterization of $F(a)$. It is well known that the Rayleigh quotient associated with (1) is given by

$$\tag{4} \mathcal{R}_a(v) = \frac{\int_\Omega (\Delta v)^2 dx + a \int_\Omega v^2 dx}{\int_\Omega |\nabla v|^2 dx},$$

where $v \in H_0^2(\Omega$. Let us first prove that $F(a)$ is concave by establishing that for any $a_0 \in [0, \infty)$ there exists $M \in I\!R$ such that

$$\tag{5} F(a) - F(a_0) \leq M(a - a_0) \quad \text{for any } a \in [0, \infty).$$

From the min–max characterization of eigenvalues (see [1, Vol. 1]) we know that

$$\tag{6} F(a) \leq \sum_{i=1}^{m} \mathcal{R}_a(v_i)$$

for every orthonormal system $\{v_1, \dots, v_m\}$ of admissible functions in $H_0^2(\Omega)$. Here $\{v_1, v_2, \dots, v_m\}$ are orthonormal with respect to $\int_\Omega \nabla v_i \nabla v_j \ dx$.

Let u_i^a denote the ith eigenfunction associated with (1), normalized to $||\nabla u_i^a||_{L^2(\Omega)} = 1$. Set $v_i = u_i^{a_0}$. Then (6) implies

$$F(a) - F(a_0) \leq (a - a_0) \sum_{i=1}^{m} \int_\Omega v_i^2 \ dx.$$

Therefore (5) holds with a positive M and $F(a)$ is strictly increasing and concave. To prove that $F(a)$ is strictly concave, suppose the contrary. Then there exists an interval $[b, c] \subset [0, \infty)$ such that

$$F(tb + (1-t)c) = tF(b) + (1-t)F(c)$$

holds for every $T \in [0, 1]$. In particular, setting $t = \frac{1}{2}$ and $a = (b+c)/2$

$$F(a) = \tfrac{1}{2}F(b) + \tfrac{1}{2}F(c). \tag{7}$$

We observe that

$$\gamma_i(b) = \mathcal{R}_b(u_i^b) \leq \mathcal{R}_b(u_i^a), \tag{8}$$

and that the same inequality holds with b replaced by c. Therefore we have from (7) and (8)

$$\begin{aligned} F(a) &\leq \frac{1}{2}\sum_{i=1}^{m} \mathcal{R}_b(u_i^a) + \frac{1}{2}\sum_{i=1}^{m} \mathcal{R}_c(u_i^a) \\ &= \sum_{i=1}^{m} \mathcal{R}_a(u_i^a) = F(a). \end{aligned}$$

But now equality must hold in (8) for every $i = 1, \ldots, m$. In particular $\gamma_1(b) = \mathcal{R}_b(u_1^a)$ and $\gamma_1(c) = \mathcal{R}_c(u_1^a)$; that is, $u = u_1^a$ is an eigenfunction corresponding to both $\gamma_1(b)$ and $\gamma_1(c)$. Subtraction of the corresponding differential equations (1) yields:

$$(\gamma_1(c) - \gamma_1(b))\Delta u + (c - b)u = 0. \tag{9}$$

If $\gamma_1(c) = \gamma_1(b)$ then (9) implies $c = b$ as desired. Otherwise u is an eigenfunction to the Laplace operator on Ω and satisfies both $u = 0$ and $\partial u/\partial n = 0$ on $\partial\Omega$, a contradiction to Hopf's second lemma. Therefore $c = b$ and this completes the proof of Theorem 1. □

Remark 1. In [10, p. 286ff] Polya and Schiffer proved concavity of sums of eigenvalues for some Neumann problems. Our result and proof are inspired by theirs. One might conjecture that each of the eigenvalues is concave in a separately. Numerical results in [5] indicate that in general this is not the case. Notice, however, that $\gamma_1(a)$ is strictly concave.

From now on we concentrate on the first eigenvalue $\gamma_1(a)$.

COROLLARY 2. *Inequalities* (2) *are strict for* $a > 0$.

Proof. By Theorem 1, equality cannot hold on the right-hand side of (2) for $a > 0$, nor on the left-hand side when $0 < a < \gamma_1^2(0)/4$. Moreover, we have after integration by parts in the denominator and by Schwarz's inequality

$$\begin{aligned} \gamma_1(a) = \mathcal{R}_a(u_1^a) &= \frac{\int_\Omega (\Delta u_1^a)^2 dx + a\int_\Omega (u_1^a)^2 dx}{\int_\Omega |\nabla u_1^a|^2 dx} \\ &\geq \left(\frac{\int_\Omega (\Delta u_1^a)^2 dx}{\int_\Omega (u_1^a)^2 dx}\right)^{1/2} + a\left(\frac{\int_\Omega (u_1^a)^2 dx}{\int_\Omega (\Delta u_1^a)^2 dx}\right)^{1/2} \geq 2\sqrt{a}. \end{aligned}$$

Thus $\gamma_1(a) = 2\sqrt{a}$ if and only if

$$\frac{\int_\Omega (\Delta u_1^a)^2 dx}{\int_\Omega (u_1^a)^2 dx} = a$$

and

$$\int_\Omega (\Delta u_1^a)^2 dx \cdot \int_\Omega (u_1^a)^2 dx = \left(\int_\Omega u_1^a \Delta u_1^a dx\right)^2,$$

and equality holds if and only if $\Delta u_1^a + \sqrt{a} u_1^a = 0$. But $u_1^a = 0 = \partial u_1^a/\partial n$ on $\partial\Omega$, so that $u_1^a = 0$. Thus, the strict inequalities

$$\max\{\gamma_1(0),\, 2\sqrt{a}\} < \gamma_1(a) < \gamma_1(0) + \frac{a}{\lambda_1} \tag{10}$$

hold. This completes the proof of Corollary 2. □

Of particular interest is the asymptotic behavior of the eigenvalue $\gamma_1(a)$ and the associated eigenfunction u_1^a as $a \to \infty$. We can give the following partial answer to this problem.

THEOREM 3. *Let u_1^a be a first eigenfunction, normalized so that $||\nabla u_1^a||_{L^2(\Omega)} = 1$, and $\gamma_1(a)$ the first eigenvalue of* (1). *Then $||u_1^a||_{L^2(\Omega)} \to 0$ and $\gamma_1(a)/\sqrt{a} \to 2$ as $a \to \infty$.*

The proof of Theorem 3 will proceed in several steps. The results of Theorems 1, 3, and inequality (2) are illustrated in Fig. 1.

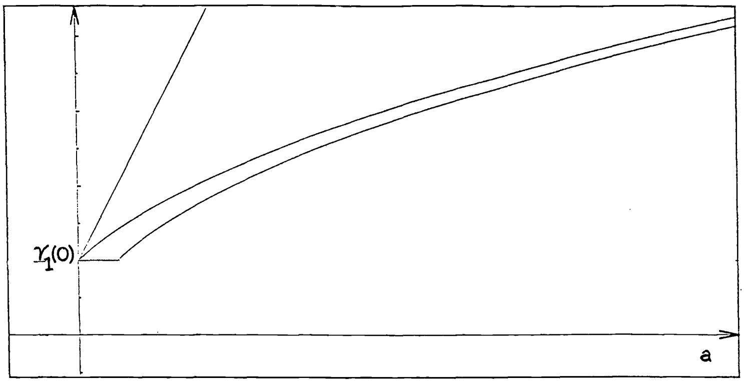


FIG. 1. $\gamma_1(a)$.

LEMMA 4. (a) *If Ω is starshaped with respect to zero, then $\gamma_1(a)/\sqrt{a}$ is decreasing, i.e.,*

$$\frac{\gamma_1(a)}{\sqrt{a}} > \frac{\gamma_1(b)}{\sqrt{b}} \quad \textit{for } 0 < a < b. \tag{11}$$

(b) *If Ω is a bounded domain, then there exists a constant $M \in (2, \infty)$ such that*

$$\frac{\gamma_1(a)}{\sqrt{a}} < M \quad \textit{as } a \to \infty. \tag{12}$$

It should be remarked that Rother [12] proved (12) under the assumptions of Lemma 4(a). To prove Lemma 4(a), recall that

$$\gamma_1(a) = \min_{u \in H_0^2(\Omega)} \mathcal{R}_a(u),$$

with $\mathcal{R}_a(u)$ defined by (4). Using the transformation

$$y_j = a^{1/4}x_j \quad \text{for } j = 1, \dots, N; \qquad v(y) = u(x(y)), \quad \frac{\partial u}{\partial x_j} = a^{1/4}\frac{\partial v}{\partial y_j},$$

the expression $\mathcal{R}_a(u)$ is converted into

$$\tilde{\mathcal{R}}_a(v) = \sqrt{a} \cdot \frac{\int_{\Omega_a}(\Delta v)^2 dx + \int_{\Omega_a} v^2 dx}{\int_{\Omega_a} |\nabla v|^2 dx},$$

where $\Omega_a = a^{1/4}\Omega$ is the image of Ω under the above transformation. Therefore $\gamma_1(a)/\sqrt{a}$ can be characterized through

$$\frac{\gamma_1(a)}{\sqrt{a}} = \min_{u \in H_0^2(\Omega_a)} \frac{\int_{\Omega_a}(\Delta v)^2 dx + \int_{\Omega_a} v^2 dx}{\int_{\Omega_a} |\nabla v|^2 dx}, \tag{13}$$

and (13) is equivalent to the original characterization of γ_1. At this point the starshapedness of Ω enters into the proof, because for starshaped domains we have

$$\Omega_a \subset \Omega_b \quad \text{for } 0 < a < b.$$

Since functions from $H_0^2(\Omega_a)$ can be continued by zero in $\overline{\Omega_b} \setminus \Omega_a$, property (11) follows from the well-known monotone dependence of eigenvalues on the domain Ω. This completes the proof of Lemma 4(a).

To prove Lemma 4(b) let B be a ball contained in Ω. Without loss of generality we may assume that Ω contains zero and that $a > 1$. Let $\tilde{u}_1^a$ be a first eigenfunction associated to the first eigenvalue $\tilde{\gamma}_1(a)$ on B. Then by the monotone dependence of γ_1 on Ω and by (11) we have

$$\frac{\gamma_1(a)}{\sqrt{a}} \leq \frac{\tilde{\gamma}_1(a)}{\sqrt{a}} \leq \tilde{\gamma}_1(1), \tag{14}$$

and this completes the proof of Lemma 4. □

Now we can prove the first statement of Theorem 3. Relation (14) implies

$$||u_1^a||^2_{L^2(\Omega)} \leq \frac{\tilde{\gamma}_1(1)}{\sqrt{a}}, \tag{15}$$

so that $u_1^a \to 0$ in $L^2(\Omega)$ of order $a^{-1/4}$. To complete the proof of Theorem 3 it suffices to combine (14) with the following result.

LEMMA 5. *If Ω is a ball or rectangular parallelepiped, then*

$$\lim_{a\to\infty} \frac{\gamma_1(a)}{\sqrt{a}} = 2. \tag{16}$$

To show (16), first in the one-dimensional case, we take $\{a_n\}_{n\in N} = \{\lambda_n\}_{n\in N} = \{n^2\pi^2\}_{n\in N}$ and $\overline{\Omega} = [0,1]$. With $\psi_n = c_n \sin(n\pi x)$, it is easy to see that there are constants $d_1, d_2 > 0$ such that for all n, with $\Omega_n = (0, 1/n) \cup (1 - 1/n, 1)$

$$\int_{\Omega_n} |\nabla\psi_n|^2 \, dx \leq \frac{d_1}{\lambda_n^{1/2}}$$

$$\int_{\Omega_n} |\psi_n|^2 \, dx \leq \frac{d_2}{\lambda_n^{3/2}},$$

where we normalize the ψ_n so that

$$\int_0^1 |\psi_n'|^2 \, dx = \lambda_n \int_0^1 \psi_n^2 \, dx = 1.$$

The eigenfunctions are uniformly oscillating on (0,1). The functions ψ_n are not admissible for the Rayleigh quotient (4), since they are not in $H_0^2(\Omega)$. In order to modify them near the boundary, we construct functions of the form $\phi_n = \psi_n \eta_\varepsilon$, where $\varepsilon = 1/n$ and η_ε is given by

$$\eta_\varepsilon(x) = \begin{cases} \phi_\varepsilon(x) & 0 \le x \le \frac{1}{n}, \\ 1 & \frac{1}{n} < x < 1 - \frac{1}{n}, \\ \phi_\varepsilon(1-x) & 1 - \frac{1}{n} \le x \le 1, \end{cases}$$

where, for $x \in (0, \varepsilon)$,

$$\phi_\varepsilon(x) = \frac{\int_0^x \, \mathrm{e}^{-\varepsilon^4/(y^2(\varepsilon^2 - y^2))} \, dy}{\int_0^\varepsilon \, \mathrm{e}^{-\varepsilon^4/(y^2(\varepsilon^2 - y^2))} \, dy}.$$

η_ε is of class C^2 in $(0,1)$, $\eta = \eta' = 0$ at $x = 0, 1$, and there are constants d_3, d_4, independent of ε such that $\max |\eta_\varepsilon'| \le d_3/\varepsilon$, $\max |\eta_\varepsilon''| \le d_4/\varepsilon^2$. Thus a tedious, but routine, calculation yields, using $\eta = \eta_{1/n}$ for notation,

$$\begin{aligned} \frac{\mathcal{R}_{\sqrt{\lambda_n}}(\psi_n \eta)}{\sqrt{\lambda_n}} &\le \frac{2 - \int_{\Omega_n} \, 2\psi_n \eta(\psi_n' \eta' + \psi_n \eta'') \, dx + \lambda_n^{-1} \int_{\Omega_n} \, (\psi_n \eta'' + \psi_n' \eta')^2 \, dx}{1 - \int_{\Omega_n} \, [(1 - \eta^2)(\psi_n')^2 + \psi_n^2 \eta \eta''] \, dx} \\ &\le \frac{2 + d_5 \lambda_n^{-1/2}}{1 - d_6 \lambda_n^{-1/2}} \end{aligned}$$

with constants d_5, d_6 independent of n. (Note that $\varepsilon = 1/n = \pi/\lambda_n^{1/2}$.) To verify (16) for arbitrary domains in higher dimensions, it is necessary to have good estimates for the local L^2 norms of the eigenfunction and its gradient near the boundary. For N-dimensional rectangles, however, the one-dimensional example is easily modified. If Ω is the unit ball in $\mathbb{R}^N$, then the radially symmetric eigenfunctions are given by

$$\psi_j(x) = c_j r^{-(N-2)/2} J_{(N-2)/2}(\sqrt{\lambda_j} r),$$

where J_ν is the usual Bessel function of order ν, the numbers $\lambda_j^{1/2}$ are the roots of J_ν in increasing order and the c_j's are normalizing constants chosen so that

$$\int_\Omega |\nabla \psi_j|^2 \, dx = \lambda_j \int_\Omega \psi_j^2 \, dx = 1,$$

or

$$\int_0^1 (\psi_j'(r))^2 r^{N-1} \, dr = \lambda_j \int_0^1 \psi_j^2(r) r^{N-1} \, dr = \omega_N^{-1}.$$

Precisely, we have

$$c_j^2 = \frac{2}{\lambda_j J_{(N-2)/2}^2(\sqrt{\lambda_j})},$$

with

$$\sqrt{\lambda_j} = (j + (N-3)/4)\pi + O\left(\frac{1}{j}\right)$$

and, for any index $\nu > -1$ as $r \to +\infty$ and some constant C_ν,

$$|J_\nu(r)| = C_\nu r^{-1/2}\left(1 + O\left(\frac{1}{r}\right)\right).$$

From these estimates we easily see that there are constants $d_1, d_2 > 0$ such that for all $j >> 1$,

$$\int_{1-1/\sqrt{\lambda_j}}^{1} \psi_j^2(r) r^{N-1}\, dr \le d_1 \lambda_j^{-3/2}$$

and

$$\int_{1-1/\sqrt{\lambda_j}}^{1} (\psi_j'(r))^2 r^{N-1}\, dr \le d_2 \lambda_j^{-1/2}.$$

Thus, if we take $\eta = \eta_\varepsilon(r)$ to be one on the ball of radius $1 - \varepsilon$, satisfy $0 < \eta < 1$ in the annular region $\{1 - \varepsilon < r < 1\}$ with $\eta(1) = \eta'(1) = 0$ and with $\varepsilon = \lambda_j^{-1/2}$, we see that with $\phi_j = \eta_\varepsilon \psi_j$, we again have

$$\frac{\mathcal{R}_{\sqrt{\lambda_j}}(\eta_\varepsilon \psi_j)}{\sqrt{\lambda_j}} \le \frac{2 + d_5 \lambda_j^{-1/2}}{1 - d_6 \lambda_j^{-1/2}}$$

for computable constants d_5, d_6. In fact we can choose η_ε so that for some d_3, d_4 the following estimates hold: $\max|\eta_\varepsilon'| \le d_3 \lambda_j^{-1/2}$ and $\max|\eta_\varepsilon''| \le d_4 \lambda_j^{-1}$. This together with (10) completes the proof of Lemma 5 and thus of Theorem 3. □

Remark 2. The limiting process in Theorem 3 can be recast as the singular perturbation problem of minimizing

$$I_\varepsilon(v) = \int_\Omega \varepsilon(\Delta v)^2 + v^2\, dx \quad \text{over} \quad \left\{ v \in H_0^2(\Omega) \mid \int_\Omega |\nabla v|^2 dx = 1 \right\}.$$

The formal limit problem for $\varepsilon = 0$ has no solution, but as the proof of Lemma 4 shows, for certain domains Ω there exists a minimizing sequence v_ε for I_0 such that $I_0(v_\varepsilon) \to 0$ as $\varepsilon \to 0$. Moreover, v_ε is highly oscillatory and the oscillations of v_ε are equidistributed. A similar qualitative behaviour has been observed by Müller in [8]. He minimized

$$\tilde{I}_\varepsilon(v) = \int_0^1 [\varepsilon(v_{xx})^2 + (v_x^2 - 1)^2 + v^2]\, dx \quad \text{over } H_0^2(0,1),$$

and showed that minimizers v_ε of $\tilde{I}_\varepsilon$ are rapidly and regularly oscillating and converge to zero in $L^2(\Omega)$. Moreover, the formal limit problem for $\varepsilon = 0$ has no solution, either. Theorem 3 shows that oscillatory behavior of this nature is not restricted to nonlinear problems, but can just as well occur for solutions of classical linear problems. In fact, physical intuition tells us that the buckled state of the plate should oscillate while its amplitude decreases as the ambient medium gets stiffer and stiffer.

Remark 3. *Linear elasticity system.* The above results were inspired by the paper [11] of W. Rother, who investigated the dependence of the first eigenvalue $\Lambda_1(a)$ of Lamé's operator on a parameter $a = (\lambda + \mu)/\mu$, where λ and μ are the Lamé constants. This eigenvalue can be characterized by

$$\Lambda_1(a) = \min\left\{ ||\nabla \underline{u}||^2 + a||\text{div}\ \underline{u}||^2 \mid \underline{u} \in [H_0^1(\Omega)]^N,\ ||\underline{u}||_{L^2(\Omega)^N} = 1 \right\}, \tag{17}$$

see, e.g., [2]. The associated system reads

$$\begin{aligned} \Delta \underline{u} + a \operatorname{grad} \operatorname{div} \underline{u} + \Lambda_1 \underline{u} &= 0 \quad \text{in } \Omega, \\ \underline{u} &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Problem (17) is related to the so-called fundamental Stokes' eigenvalue:

$$m_1 = \min \{ ||\nabla \underline{u}||^2 \mid \underline{u} \in [H_0^1(\Omega)]^N, \ \operatorname{div} \underline{u} = 0, \ ||\underline{u}||_{L^2(\Omega)^N} = 1 \}. \tag{18}$$

It was shown in [2] that m_1 is an upper bound for $\Lambda_1(a)$. In [11] Rother showed that $\Lambda_1(a)$ is increasing in a. Using the ideas above it can easily be shown that in fact $\sum_{i=1}^m \Lambda_i(a)$ is concave in a. The lower bound

$$\sum_{i=1}^{m} \Lambda_i(a) \geq \frac{3}{5} \left(\frac{2\pi^2}{V} \right)^{2/3} m^{5/3}$$

was derived in [6]. Under the technical assumption that $\partial\Omega \in C^{0,1}$, Rother showed that the upper bound m_1 is optimal in the sense that

$$\lim_{a \to \infty} \Lambda_1(a) = m_1. \tag{19}$$

In [4], it was shown that

$$\lambda_1 \leq \Lambda_1(a) \left(1 + \frac{a}{3}\right) \lambda_1.$$

See Fig. 2 for a graphical summary of the discussion of the results for $\Lambda_1(\cdot)$.

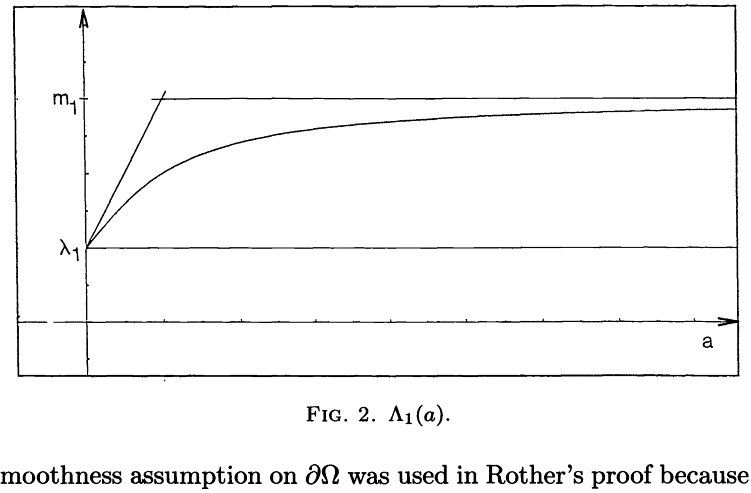


FIG. 2. $\Lambda_1(a)$.

The smoothness assumption on $\partial\Omega$ was used in Rother's proof because he decomposed the eigenfunctions orthogonally into divergence free and remaining components, and he then applied some results for the divergence operator. We can avoid these difficulties (and thus derive (19) without any regularity assumption on $\partial\Omega$) as follows: let $\underline{u}_n$ be a sequence of eigenvectors associated with the eigenvalue $\Lambda_1(n)$ and suppose that $n \to \infty$. Since $\Lambda_1(n) \leq m_1$ we know that $\underline{u}_n$ is uniformly bounded in $[H_0^1(\Omega)]^N$ and that div $\underline{u}_n \to 0$ as $n \to \infty$. Therefore, after possibly passing to a subsequence,

$\underline{u}_n$ has a weak limit $\underline{u}_\infty$ in $[H_0^1(\Omega)]^N$. Moreover, $\underline{u}_\infty$ has unit length in $L^2(\Omega)^N$. Since the map $v \mapsto \int_\Omega (\text{div } v)^2\, dx$ is convex and hence weakly lower semicontinuous in $[H_0^1(\Omega)]^N$, we conclude that $\text{div}\, \underline{u}_\infty = 0$. Finally it should be noted that u_n converges strongly in $[H_0^1(\Omega)]^N$ to u_∞, since $||u_n|| \to ||u_\infty||$ and u_n converges weakly. Therefore (19) must hold.

Remark 4. *Clamped plate under tension.* Instead of (1) consider the eigenvalue problem

$$\begin{aligned} \Delta\Delta u - a\Delta u - \Gamma u &= 0 \quad \text{in } \Omega, \\ u = \frac{\partial u}{\partial n} &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{20}$$

where $a = T/D$ is given. T describes tension and D the flexural rigidity of the plate. The eigenvalues Γ_i of (20) are characterized by means of the following Rayleigh quotient on $H_0^2(\Omega)$

$$\mathcal{R}_a(v) = \frac{\int_\Omega \{(\Delta v)^2 + a|\nabla v|^2\}\, dx}{\int_\Omega v^2\, dx}. \tag{21}$$

As in the proof of Theorem 1 it can be shown that $\sum_{i=1}^m \Gamma_i(a)$ is strictly concave and strictly increasing in a. The following estimates hold for $\Gamma_1(a)$:

$$\Gamma_1(0) + a\lambda_1 \le \Gamma_1(a) \le \Gamma_1(0) + a\sqrt{\Gamma_1(0)} \quad \text{for } a > 0; \tag{22}$$

see [9]. Here again λ_1 is the lowest eigenvalue of the corresponding fixed membrane problem and $\sqrt{\Gamma_1(0)}$ is the fundamental frequency of a clamped plate in the absence of tension. Notice that (22) is sharp for $a = 0$, and that (22) implies that the curve $(a, \Gamma_1(a))$ stays inside a certain cone. A consequence of our results is that the inequalities in (22) are necessarily strict, an assertion not claimed by Payne. The numerical results of [14] indicate that $\Gamma_1(a)$ is a concave function with an asymptote whose slope is not smaller than λ_1, see Figs. 3 and 4.

We claim that the eigenfunction u_a associated to $\Gamma_1(a)$ converges to ψ_1 and $\Gamma_1(a)$ converges to λ_1 as $a \to \infty$; see Fig. 4. Indeed, $\Gamma_i(a) \to \lambda_i$ for $i = 1, 2, \dots$, as $a \to \infty$. This is easy to see once we realize that letting $a \to \infty$ is equivalent to letting the flexural rigidity of the plate tend to zero. Thus, in the limit the plate should behave like a membrane. Setting $\varepsilon = 1/a$ we can rewrite the differential equation in (20) as

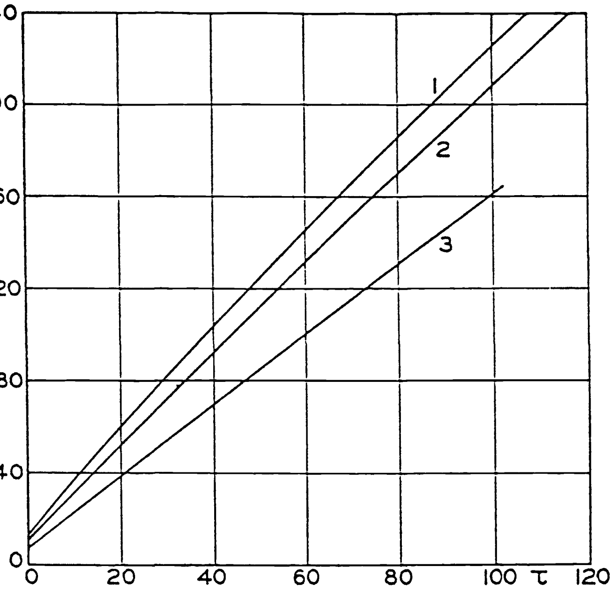
$$\varepsilon\Delta\Delta u - \Delta u - \lambda(\varepsilon)u = 0,$$

and view this differential equation as a singular perturbation of the membrane equation given at the end of the Introduction. In fact, asymptotic expansions for $\lambda_i(\varepsilon)$ and its corresponding eigenfunctions are well known and recorded, for instance, in [2, p. 392], [3], [15].

2. More general results, the linear case. The above result can be generalized in several ways: For example, let H be a Hilbert space and D_1, D_2 ($D_2 \subset D_1 \subset H$) be dense linear subspaces on which the nonnegative, selfadjoint operators E_1, E_2 are defined respectively. We shall assume that E_2 is strictly positive, i.e., $(x, E_2x) > 0$ unless $x = 0$. Let $(\cdot,\cdot)$ and $||\ \ ||$ denote the scalar product and corresponding norm on H. We let $\overline{D_i}$ be the completion of D_i in the norm $\{(x, E_ix) + ||x||^2\}^{\frac{1}{2}}$.

(A.1) For every $a \ge 0$ there exists $y_a \in \overline{D_2}$ such that

$$\begin{aligned} \Gamma(a) &= \inf\{(x, E_2x) + a(x, E_1x) \mid ||x|| = 1, x \in \overline{D_2}\ \} \\ &= (y_a, E_2y_a) + a(y_a, E_1y_a) =: J_a(y_a). \end{aligned}$$



Curve 1: Clamped Square Plate ($\pi/2 \geqq x, y \geqq -\pi/2$)
Curve 2: Clamped Circular Plate ($r = \sqrt{\pi}$)
Curve 3: Clamped Circular Plate ($r = 2$)

FIG. 3. *Numerical approximations of* $\Gamma_1(a)$, *copied from* [14] ($\tau = a$).

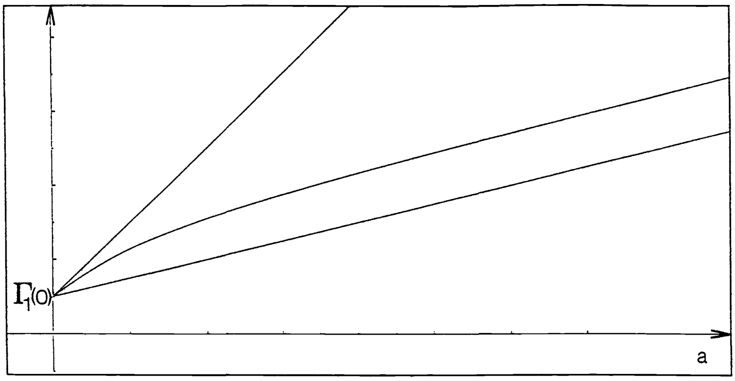


FIG. 4. $\Gamma_1(a)$.

THEOREM 6. *Suppose that* (A.1) *holds.*

(i) *If* $a < b$, *then*

$$(y_b, E_1 y_b) \leq \frac{\Gamma(b) - \Gamma(a)}{b - a} \leq (y_a, E_1 y_a), \tag{23}$$

and $\Gamma(a)$ *is a monotone nondecreasing function of* a.

(ii) $\Gamma(a)$ *is a concave function of* a.

(iii) $\Gamma(a)$ *is strictly increasing on an interval* $(\alpha, \beta) \subset (0, \infty)$ *if and only if* $(y_a, E_1 y_a) > 0$ *for all* $a \in (\alpha, \beta)$.

(iv) $\Gamma(a)$ *is strictly concave if* E_1 *and* E_2 *have no common eigenvector.*

COROLLARY 7. (i) $|\Gamma(b) - \Gamma(a)| \leq (y_0, E_1 y_0)\,|b - a|$. ($\Gamma(a)$ *is Lipschitz continuous.*)

(ii) $\tilde{F}(a) = (y_a, E_1 y_a)$ *is a nonincreasing function of* a.

(iii) $\Gamma(0) + a \lim_{b\to\infty}(y_b, E_1 y_b) \leq \Gamma(a) \leq \Gamma(0) + a(y_0, E_1 y_0)$.

The proof of Corollary 7 is immediate. Let us prove Theorem 6. Statement (i) is straightforward, since $\Gamma(b) = J_b(y_b) \leq J_b(y_a)$ and $\Gamma(a) = J_a(y_a) \leq J_a(y_b)$. To prove concavity, and thus statement (ii), we have $t\Gamma(a) = tJ_a(u_a) \leq tJ_a(y_t)$ and $(1-t)\Gamma(b) = (1-t)J_b(y_b) \leq (1-t)J_b(y_t)$ for any $t \in (0,1)$, where y_t is shorthand for $y_{ta+(1-t)b}$. Adding these inequalities, we have

$$t\Gamma(a) + (1-t)\Gamma(b) \leq \Gamma(ta + (1-t)b). \tag{24}$$

In order to prove (iii) we need only observe from (23) that $(y_a, E_1 y_a) > 0$ in (α, β) if and only if $\Gamma(a)$ is strictly increasing in (α, β). To prove (iv) notice that equality holds in (24) if and only if $\Gamma(b) = J_b(y_b) = J_b(y_t)$ and $J_a(y_t) = J_a(y_a) = \Gamma(a)$. The latter is equivalent to

$$E_2 y_t + bE_1 y_t - \Gamma(b) y_t = 0 \tag{25}$$

and

$$E_2 y_t + aE_1 y_t - \Gamma(a) y_t = 0.$$

Upon subtraction we see that

$$E_1 y_t = \frac{\Gamma(b) - \Gamma(a)}{b-a} y_t, \tag{26}$$

so that y_t is an eigenvector of the operator E_1. But now, using (25) or (26) it can be seen that y_t is an eigenvector of E_2, too. This proves Theorem 6. □

Remark 5. One particular consequence of (23) or Corollary 7(ii) is the following: If $(y_a, E_1 y_a) = 0$ for some $a > 0$, then $\Gamma(a)$ is constant on $[a, \infty)$ and $(y_b, E_1 y_b) = 0$ for $b \in [a, \infty)$.

In order to obtain information on the limit $a \to \infty$ we need more assumptions about the relationship between E_1 and E_2, e.g., the following assumption.

(A.2) There exists $\alpha \in (0,1]$ such that $(x, E_1 x) \leq (x, E_2 x)^\alpha$ for all $x \in \overline{D_2}$, $||x|| = 1$.

Then

$$(y_a, E_1 y_a) \leq (y_0, E_1 y_0) \leq (y_0, E_2 y_0)^\alpha = \Gamma(0)^\alpha. \tag{27}$$

For example, for the plate under tension, $E_2 u = \Delta\Delta u$, $E_1 u = \Delta u$ on $H_0^2(\Omega)$ and $H_0^1(\Omega)$ respectively, property (A.2) holds with $\alpha = \frac{1}{2}$. Or for the Lamé operator $E_2 = \Delta$ and $E_1 = \nabla(\text{div }) $ on $[H_0^1(\Omega)]^n$ property (A.2) holds with $\alpha = 1$. Also, by unique continuation the hypothesis of Theorem 6(iv) holds for this example.

$$\Gamma(a) \leq \inf\ \{(x, E_2 x) \mid ||x|| = 1,\ x \in \overline{D_2},\ (x, E_1 x) = 0\}. \tag{28}$$

One has to distinguish two cases: (1) The infimum in (28) is taken over an empty set. (2) The infimum in (28) is taken over a nonempty set.

In both cases the family $\{(y_a, E_1 y_a)\}_{\{a>0\}}$ is bounded in view of (27). In the second case, however $(x, E_1 x)^{1/2}$ is only a seminorm, since there are vectors for which $(x, E_1 x) = 0$ and consequently the kernel $\operatorname{Ker} E_1 = \{x \mid E_1 x = 0\}$ is not trivial. If the infimum in (28) is taken over an empty set, it is ∞ by convention and we assume the following.

(A.3.1) If $\operatorname{Ker} E_1 = \{0\}$, then $\overline{D_1} = \overline{D_2}$ and sequence which is bounded in $\overline{D_1}$ posesses a subsequence which converges strongly in H and weakly in $\overline{D_1}$. Equivalently, $\overline{D_1}$ is compactly embedded in H.

If the infimum is taken over a nonempty set, $\Gamma(a) \leq M < \infty$ for all a and some M. In that case we assume (A.3.2)

(A.3.2) If $\operatorname{Ker} E_1 \neq \{0\}$, then every sequence which is bounded in $\overline{D_2}$ possesses a subsequence which converges strongly in H and weakly in $\overline{D_1}$.

For reasons that will become obvious in the proof of Theorem 8(iii), we need an additional assumption, namely, (A.4).

(A.4) Let $\mathcal{E}_a = \{y_a \mid J_a(y_a) = \Gamma(a)\}$. For every $a \geq 0$ there exists $\tilde{y}_a \in \mathcal{E}_a$ such that

$$(\tilde{y}_a, E_1\tilde{y}_a) = \inf\{(y_a, E_1 y_a) \mid y_a \in \mathcal{E}_a\} =: F(a).$$

We can now establish an analogue to (19) or Remark 4.

THEOREM 8. *Suppose that* (A.1), (A.2), *and* (A.3) *hold. Then*

(i) $\lambda_1 = \lim_{a\to\infty} (y_a, E_1 y_a)$ *is the slope of the linear asymptote of* $\Gamma(a)$ *and* $\lambda_1 \leq (y_0, E_1 y_0) \leq \Gamma(0)^\alpha$.

(ii) *Moreover* λ_1 *is the smallest eigenvalue of* E_1 *and the family* $\{u_a\}$ *contains a sequence* $\{u_{a_n}\}$ *which converges strongly in* H *and weakly in* $\overline{D_1}$ *to an element of the first eigenspace of* E_1 *as* $a_n \to \infty$.

(iii) *Whenever* $\Gamma'(a)$ *exists and* (A.4) *holds, then* $\Gamma'(a) = \inf\{(y_a, E_1 y_a) \mid y_a \in \mathcal{E}_a\}$.

The proof of (i) follows from (27). To prove (ii) we notice that (i) and (A.3) imply the existence of a sequence y_{a_n} which converges weakly in $\overline{D_1}$ and strongly in H to a limit as $a_n \to \infty$ we distinguish the above two cases.

(1) If $\operatorname{Ker} E_1 = \{0\}$ the set $\{(y_a, E_1 y_a)\}$ is uniformly bounded and $\{y_a\}$ possesses a sequence which converges strongly in H and weakly in $\overline{D_i}$ to an element of $\overline{D_i}$.

(2) If $\operatorname{Ker} E_1 \neq \{0\}$ the set $\{(y_a, E_2 y_a)\}$ is uniformly bounded and $\{y_a\}$ possesses a sequence which converges strongly in H and weakly in $\overline{D_1}$ to an element of $\overline{D_1}$. Let us call the limit element y_∞. We have

$$\frac{(y_a, E_2\phi)}{a} + (y_a, E_1\phi) - \frac{\Gamma(a)}{a}(y_a, \phi) = 0$$

for every $\phi \in \overline{D_2}$, so that y_∞ is an eigenfunction for E_1:

$$(y_\infty, E_1\phi) - m(y_\infty, \phi) = 0$$

with $d = \lim_{a\to\infty} \Gamma(a)/a$. Notice that $||y_\infty|| = 1$ by assumption (A.3). By definition of λ_1, $d \geq \lambda_1$. It remains to show that $d = \lambda_1$, and we suppose in contrast that $d > \lambda_1$. Clearly E_1 has a smallest (nonnegative) eigenvalue $\mu_1 \leq \lambda_1$ and some associated eigenfunction $\psi_1 \in \overline{D_1}$. We claim $d = \lambda_1 = \mu_1$. Let $\varepsilon < d - \mu_1$. Since D_2 is dense in H and $D_2 \subset D_1 \subset H$, we can approximate $\psi_1 \in \overline{D_1}$ with a function $\phi_\varepsilon \in D_2$ such that $||\phi_\varepsilon|| = 1$ and $(\phi_\varepsilon, E_1\phi_\varepsilon) < \mu_1 + \varepsilon$. But this contradicts the choice of ε because

$$d \leq \lim_{a_n\to\infty} \frac{1}{a_n}(\phi_\varepsilon, E_2\phi_\varepsilon) + (\phi_\varepsilon, E_1\phi_\varepsilon) \leq \mu_1 + \varepsilon,$$

and thus concludes the proof of (ii). To prove (iii) we assume (A.4). Then for any decreasing sequence $a_n \to a$ there exists a number M such that $J_a(\tilde{y}_a) \leq J_{a_n}(\tilde{y}_{a_n}) \leq$

$J_{a_1}(\tilde{y}_{a_1}) \leq M$. Therefore $\{\tilde{y}_{a_n}\}$ has a subsequence, still denoted by $\{\tilde{y}_{a_n}\}$ with a limit $\tilde{y}_{a_\infty}$. We obtain

$$\Gamma(a) \leq J_a(\tilde{y}_{a_\infty}) \leq \liminf \Gamma(a_n) \leq \Gamma(a)$$

from the definition and the continuity properties of Γ. Therefore $y_{a_\infty} \in \mathcal{E}_a$. Furthermore, due to (27) and the monotonicity of F,

$$F(a) \leq (y_{a_\infty}, E_1 y_{a_\infty}) \leq \liminf(\tilde{y}_{a_n}, E_1\tilde{y}_{a_n}) = \liminf F(a_n) \leq F(a).$$

This proves that $F(a)$ equals the one-side derivative of Γ from the right. Since $\Gamma'(a)$ is assumed to exist, the proof of Theorem 8 is complete. □

3. More general results, the nonlinear case. For $i = 0, 1, 2$, let $J_i : X_i \to \mathbb{R}^+$ be a nonnegative weakly lower semicontinuous functional on a separable Banach space X_i.

THEOREM 9. *Suppose that $X_2 \subset X_1 \subset X_0$, and that there exists a unique minimizer u_0 in X_2 of $J_1(v)$ in $X_1 \cap \{v \mid J_0(v) = 1\}$. Let u_ε be a minimizer of $J_\varepsilon(v) := \varepsilon J_2(v) + J_1(v)$ on $X_2 \cap \{v \mid J_0(v) = 1\}$.*

(i) *Then $\Gamma(\varepsilon) = J_\varepsilon(u_\varepsilon)$ is monotone nondecreasing and concave in ε.*

(ii) *If X_1 is compactly embedded in X_0 and if J_1 is coercive, then u_ε converges to u_0 weakly in X_1 and strongly in X_0.*

(iii) *If X_2 is compactly embedded in X_1, then u_ε converges to u_0 weakly in X_2 and strongly in X_1 and X_0.*

The proof is straightforward if we use ideas from the proofs of Theorems 1 and 8. As an application for Theorem 9 consider the eigenvalue problem

$$\begin{aligned} \varepsilon\Delta\Delta u - \text{div}\,(|\nabla u|^{p-2}\nabla u) - \Gamma|u|^{p-2}u &= 0 \quad \text{in } \Omega \subset \mathbb{R}^n, \\ u = \frac{\partial u}{\partial n} &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{29}$$

for $1 < p < 2n/(n-p)$. Here $J_2(v) = ||\Delta v||^2_{L^2(\Omega)}$, $J_1(v) = ||\nabla v||^2_{L^p(\Omega)}$ and $J_0(v) = ||v||_{L^p(\Omega)}$, while $X_2 = H^2_0(\Omega)$, $X_1 = W^{1,p}_0(\Omega)$ and $X_0 = L^p(\Omega)$. Then, as $\varepsilon \to 0$, the solutions of (29) converge to the (unique) ground state of the formal limit problem,

$$\begin{aligned} \text{div}\,(|\nabla u|^{p-2}\nabla u) + \lambda|u|^{p-2}u &= 0 \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

For more details on this eigenvalue problem see, e.g., [7], [13].

Notes added in proof. Professor F. Goerisch has kindly informed us of [16], in which it is shown that the entire spectrum of the elasticity operator converges to the spectrum of the Stokes operator. Therefore Remark 3 of this paper extends to all eigenvalues.

In [17], the author re-establishes the results of [11] in three dimensions. The author's method of proof relies on the decomposition of the Lamé operator using quaternians and a generalized Cauchy–Riemann operator. His result thus appears to be restricted to three dimensions. However, no regularity of the boundary is required.

Acknowledgment. We thank L. Frank for bringing references [2], [3], [15] to our attention.

REFERENCES

[1] R. Courant and D. Hilbert, *Methoden der Mathematischen Physik*, Springer-Verlag, Heidelberg, 1968.

[2] L. S. Frank, *Coercive singular perturbations, eigenvalue problems and bifurcation phenomena*, Ann. Mat. Pura Appl. (4), 148 (1987), pp. 367–395.

[3] P. P. N. de Groen, *Singular perturbation of spectra*, in Asymptotic Analysis, F. Verhulst, ed., Lecture Notes in Math. 711, Springer-Verlag, New York, 1979, pp. 9–32.

[4] B. Kawohl and G. Sweers, *Remarks on eigenvalues and eigenfunctions of a special elliptic system*, Z. Angew. Math. Phys., 38 (1987), pp. 730–740.

[5] A. W. Leissa, *On a curve veering aberration*, Z. Angew. Math. Phys. (J. Appl. Math. Phys.), 25 (1974), pp. 99–111.

[6] H. A. Levine and M. H. Protter, *Unrestricted lower bounds for eigenvalues for classes of elliptic equations and systems of equations with applications in elasticity*, Math. Methods Appl. Sci., 7 (1985), pp. 210–222.

[7] P. Lindquist, *On the equation* $div(|\nabla u|^{p-2}\nabla u) + \lambda |u|^{p-2}u = 0$, Proc. Amer. Math. Soc., 109 (1990), pp. 157–164.

[8] S. Müller, *Minimizing sequences for nonconvex functionals, phase transitions and singular perturbations*, in Problems Involving Change of Type, K. Kirchgässner, ed., Lecture Notes in Phys. 359, Springer-Verlag, New York, 1990, pp. 31–44.

[9] L. E. Payne, *New isoperimetric inequalities for eigenvalues and other physical quantities*, Comm. Pure Appl. Math., 9 (1956), pp. 531–542.

[10] G. Polya and M. Schiffer, *Convexity of functionals by transplantation*, J. Anal. Math., 3 (1953/54), pp. 245–345.

[11] W. Rother, *New estimates for the first eigenvalue of Lamé's operator*, Z. Angew. Math. Mech., 69 (1989), pp. 451–452.

[12] ———, *New bounds for the first eigenvalue of an elliptic equation occurring in the buckling problem for the plate*, Appl. Anal., 31 (1988), pp. 57–61.

[13] S. Sakaguchi, *Concavity properties of solutions to some degenerate quasilinear elliptic Dirichlet problems*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 14 (1987), pp. 401–421.

[14] A. Weinstein and Wei Zang Chien, *On the vibrations of a clamped plate under tension*, Quart. Appl. Math., 1, Amer. Math. Soc., Providence, RI, 1943, pp. 61–68.

[15] M. I. Vishik and L. A. Ljusternik, *The solution of some perturbation problems of matrices and selfadjoint or non-selfadjoint differential equations*, Russian Math. Surveys, 15, pp. 1–73.

[16] L. A. Minin, *ε-Approximation of the spectrum of the Stokes problem*, Soviet Math. Dokl., 32 (1985), pp. 211–214.

[17] K. Gurlebeck, *Lower and upper bounds for the first eigenvalue for the Lamé system*, in Boundary value and initial value problems in complex analysis: studies in complex analysis and its applications to partial differential equations, 1 (Halle, 1988), pp. 184–192, Pitman Res. Notes Math. Ser., 256, Longman Sci. Tech., Harlow, 1991.

SIAM J. MATH. ANAL.
Vol. 24, No. 2, pp. 341-364, March 1993

REGULARITY OF BOUNDARIES OF QUADRATURE DOMAINS IN TWO DIMENSIONS*

MAKOTO SAKAI†

Abstract. First, the author discusses regularity of the boundary of a bounded quadrature domain. It is shown that the boundary is very beautiful and it consists of regular, degenerate, double, and cusp points. Next the author discusses an unbounded quadrature domain and defines its global Schwarz function, showing that it is obtained as a translation of the inversion of a bounded quadrature domain if it is not dense in the whole plane. The author gives a complete description of unbounded quadrature domains of point differential functionals of finite order and determines all null quadrature domains in two dimensions.

Key words. quadrature domains, Schwarz functions, inversions

AMS(MOS) subject classifications. 30E20, 30E05, 30E99

In a previous paper [11], the author proved a regularity theorem on a boundary having a Schwarz function. Let Ω be an open subset of a disk $B_r(\zeta_0)$ with radius r and center ζ_0 such that the boundary $\partial\Omega$ of Ω contains the center ζ_0, and let $\Gamma = (\partial\Omega) \cap B_r(\zeta_0)$. We call a function S defined on $\Omega \cup \Gamma$ a Schwarz function of $\Omega \cup \Gamma$, more precisely, a Schwarz function of $\Omega \cup \Gamma$ at ζ_0 or in $B_r(\zeta_0)$, if

(i) S is holomorphic in Ω;

(ii) S is continuous on $\Omega \cup \Gamma$;

(iii) $S(\zeta) = \bar{\zeta}$ on Γ, where $\bar{\zeta}$ denotes the complex conjugate of ζ.

REGULARITY THEOREM [11]. *If there is a Schwarz function of* $\Omega \cup \Gamma$ *at* ζ_0, *then* ζ_0 *is* (1) *a regular,* (2) *degenerate,* (3) *double, or* (4) *cusp point of* Γ. *Namely, there is a small disk* $B = B_\delta(\zeta_0)$, *and one of the following must occur:*

(1) $\Omega \cap B$ *is simply connected and* $\Gamma \cap B$ *is a regular real analytic simple arc passing through* ζ_0.

(2) $\Gamma \cap B = \{\zeta_0\}$ *or* $\Gamma \cap B$ *is an infinite set accumulating at* ζ_0, *and is contained in a uniquely determined regular real analytic simple arc passing through* ζ_0. $\Gamma \cap B$ *is a proper subset of the arc or the whole arc.* $\Omega \cap B$ *is equal to* $B \setminus \Gamma$.

(3) $\Omega \cap B$ *consists of two simply connected components,* Ω_1 *and* Ω_2. $(\partial\Omega_1) \cap B$ *and* $(\partial\Omega_2) \cap B$ *are distinct regular real analytic simple arcs passing through* ζ_0. *They are tangent to each other at* ζ_0.

(4) $\Omega \cap B$ *is simply connected, and* $\Gamma \cap B$ *is a regular real analytic simple arc except for a cusp at* ζ_0. *The cusp is pointing into* $\Omega \cap B$. *It is a very special one. There is a holomorphic function* T *defined on a closed disk* $\overline{B_\varepsilon(0)}$ *such that*

(i) $T(0) = \zeta_0$, $T'(0) = 0$ *and* $T''(0) \neq 0$;

(ii) T *is univalent on the closure* $\bar{H}$ *of a half disk* $H = \{w \in B_\varepsilon(0); \operatorname{Im} w > 0\}$;

(iii) T *satisfies* $\Gamma \cap B \subset T((-\varepsilon, \varepsilon))$ *and* $T(\bar{H}) \subset \Omega \cup \Gamma$, *where* $(-\varepsilon, \varepsilon) = \{w; -\varepsilon < w = \operatorname{Re} w < \varepsilon\}$.

Conversely, if one of statements (1)-(4) holds, then $(\Omega \cap B_\rho(\zeta_0)) \cup (\Gamma \cap B_\rho(\zeta_0))$ has a Schwarz function for some $\rho > 0$.

In this paper, we apply the Regularity Theorem to quadrature domains and show regularity of boundaries of quadrature domains in two dimensions. Let μ be a complex measure on the complex plane $\mathbb{C}$. A nonempty open set Ω in $\mathbb{C}$ is called a quadrature

* Received by the editors December 17, 1991; accepted for publication (in revised form) July 23, 1992.

† Department of Mathematics, Tokyo Metropolitan University, Minami-Ohsawa 1-1, Hachioji-shi, Tokyo, 192-03, Japan.

domain of μ if $|\mu|(\mathbb{C}\backslash\Omega)=0$ and if $\int|f|d|\mu|<+\infty$ and

$$\int f\,d\mu=\iint_{\Omega} f(z)\,dx\,dy \qquad (z=x+iy)$$

for every holomorphic and integrable function f in Ω.

Quadrature domains are closely related to domains that are solutions of an inverse problem in potential theory. The Newtonian potential of a measure μ is equal to that of the quadrature domain of μ in the exterior of the quadrature domain. Quadrature domains are also closely related to Hele-Shaw flows with a free boundary. We interpret a quadrature domain as a solution of a Hele-Shaw flow free boundary problem. On the other hand, we can express the solution of the Hele-Shaw flow free boundary problem as a quadrature domain of some measure. For these and further applications, we refer the reader to [9].

This paper consists of two parts. We devote the first part to the study of regularity of boundaries of bounded quadrature domains. In the second part we give a broad perspective of unbounded quadrature domains.

If a quadrature domain Ω of μ is bounded, then $1/(z-\zeta)$ is holomorphic and integrable on Ω for every fixed $\zeta\in\mathbb{C}\backslash\Omega$. Hence the Cauchy transform $\hat{\Omega}(\zeta)=\int_{\Omega}1/(z-\zeta)\,dx\,dy$ of Ω is equal to the Cauchy transform $\hat{\mu}(\zeta)=\int 1/(z-\zeta)\,d\mu(z)$ of μ on $\mathbb{C}\backslash\Omega$. Since $\hat{\Omega}(z)+\pi\bar{z}$ is holomorphic in Ω, we see that

$$S(z)=\frac{\hat{\Omega}(z)+\pi\bar{z}-\hat{\mu}(z)}{\pi}$$

is a Schwarz function of $(\Omega\cap B)\cup((\partial\Omega)\cap B)$ if B and the support of μ are disjoint, where $B=B_r(\zeta_0)$ and $\zeta_0\in\partial\Omega$. Applying our Regularity Theorem, we obtain a regularity theorem on boundaries of bounded quadrature domains. Let Ω be a bounded quadrature domain of μ such that the support of μ is contained in Ω. If we make a new domain $[\Omega]$ by adding all degenerate boundary points of Ω to Ω, then $[\Omega]$ is also a quadrature domain of μ, and the boundary of $[\Omega]$ consists of a finite number of real analytic closed curves having at most a finite number of double and cusp points in the sense of the Regularity Theorem.

In contrast with many known results on bounded quadrature domains, only a few results are known on unbounded quadrature domains. For example, let δ_0 be the Dirac measure at the origin. Then the bounded quadrature domain of δ_0 is the simplest quadrature domain. It is determined uniquely and is equal to a disk with center at the origin and radius $1/\sqrt{\pi}$. We know that there are many unbounded quadrature domains of δ_0, but we do not know all of them yet.

Here is a program proposed by Shapiro [12]. Let Ω be an unbounded quadrature domain of a complex measure with compact support. Let S be a Schwarz function defined on $\bar{\Omega}\backslash\overline{B_R}$ for some large disk $B_R=B_R(0)$. Namely, let S be a function such that (i) S is holomorphic in $\Omega\backslash\bar{B}_R$, (ii) S is continuous on $\bar{\Omega}\backslash\overline{B_R}$, (iii) $S(\zeta)=\bar{\zeta}$ on $(\partial\Omega)\backslash\overline{B_R}$, and (iv) $|S(z)|\leqq\alpha|z|$ on $\Omega\backslash\overline{B_R}$ for some constant $\alpha>0$. We can show the existence of the Schwarz function by using the generalized Cauchy transform. If

$$(*)\qquad S(z)\to\infty \quad \text{as } z\in\Omega\to\infty,$$

then $S_i(z)=1/S(1/z)$ is holomorphic in $\Omega'=\{1/z;\ z\in\Omega\}\cap B_r(0)$ for some r. By setting $S_i(0)=0$, we see that S_i is continuous on $\Omega'\cup\Gamma'$ and satisfies $S_i(\zeta)=\bar{\zeta}$ on Γ', where $\Gamma'=(\partial\Omega')\cap B_r(0)$. Thus S_i is a Schwarz function of $\Omega'\cup\Gamma'$ at 0. Shapiro [12] showed that if $\partial\Omega$ satisfies some regularity hypotheses, then $(*)$ is satisfied and the unbounded quadrature domain is obtained as an inversion of a bounded quadrature domain.

We shall show, by applying the Fuchs theorem, that (*) is actually satisfied if $\partial\Omega$ is unbounded. This will enable us to carry out the Shapiro program on unbounded quadrature domains. In particular, we obtain a complete description of unbounded quadrature domains of finite order.

During the first reading of this paper, one could omit § 2, except for the definition of the Schwarz function of an unbounded set and the statement of Corollary 2.6. Throughout the paper, $\bar{z}$ denotes the complex conjugate of z if z is a complex number. For a subset E of $\mathbb{C}$, $\bar{E}$ denotes the closure of E. We usually denote a boundary point by ζ instead of z.

I. Bounded quadrature domains.

1. Bounded quadrature domains of a complex measure, which contain the support of the measure. In the introduction we have already mentioned regularity of boundaries of bounded quadrature domains. We summarize it as the following proposition.

PROPOSITION 1.1. *Let Ω be a bounded quadrature domain of a complex measure μ, and let $\zeta_0 \in (\partial\Omega)\backslash\operatorname{supp}\mu$. Then there is a disk $B = B_\delta(\zeta_0)$, and one of statements (1)-(4) of the Regularity Theorem holds.*

COROLLARY 1.2. *Let Ω be a bounded quadrature domain of a complex measure μ. If $\zeta \in \partial\Omega$ is neither a regular, degenerate, double, nor a cusp point of $\partial\Omega$ in the sense of our Regularity Theorem, then $\zeta \in \operatorname{supp}\mu$, namely, $|\mu|(B_r(\zeta)) > 0$ for every $r > 0$.*

For example, let μ be a positive measure defined by $d\mu = 2(1-\sqrt{|x|})\,dx$ on $(-1, 1)$, and let $\Omega = \{z = x+iy;\ -1 < x < 1,\ x^2/2 - \frac{1}{2} < y < -x^2/2 + \frac{1}{2}\}$. Then Ω is the unique bounded quadrature domain of μ; see [9, Ex. 14.10, p. 123]. $(\partial\Omega)\backslash\{-1, 1\}$ consists of two real analytic simple arcs, but -1 and 1 are two points on $\partial\Omega$ that are neither a regular, degenerate, double, nor a cusp point. In this case, $\operatorname{supp}\mu = [-1, 1]$ and $\operatorname{supp}\mu$ contains -1 and 1, as asserted in Corollary 1.2.

Now we discuss a bounded quadrature domain Ω of a complex measure μ such that $\operatorname{supp}\mu$ is contained in Ω. To make things clear, we introduce a global Schwarz function of a bounded open set.

DEFINITION 1.3. Let Ω be a nonempty bounded open set in $\mathbb{C}$. A function S defined on $\bar{\Omega}\backslash K$ for some compact subset K of Ω is called a global Schwarz function of Ω, or a global Schwarz function of Ω holomorphic in $\Omega\backslash K$, if

(i) S is holomorphic in $\Omega\backslash K$;
(ii) S is continuous on $\bar{\Omega}\backslash K$;
(iii) $S(\zeta) = \bar{\zeta}$ on $\partial\Omega$.

The global Schwarz function of a bounded open set Ω holomorphic in $\Omega\backslash K$ is determined uniquely if $\Omega\backslash K$ is connected; see the note after Definition 3.1 in [11]. Next we prepare the following definition and proposition.

DEFINITION 1.4. Let Ω be an open proper subset of a disk $B_r(\zeta_0)$. Let g be a function continuous on $B_r(\zeta_0)\backslash\Omega$ and holomorphic in the interior of $B_r(\zeta_0)\backslash\Omega$. We say that g can be extended holomorphically from $B_r(\zeta_0)\backslash\Omega$ onto $B_r(\zeta_0)$ if there is a holomorphic function f in $B_r(\zeta_0)$ such that $f(z) = g(z)$ on $B_r(\zeta_0)\backslash\Omega$.

PROPOSITION 1.5. *Let Ω be an open subset of $B_r(\zeta_0)$ such that $\zeta_0 \in \partial\Omega$, and let $\Gamma = B_r(\zeta_0) \cap \partial\Omega$. Then $\Omega \cup \Gamma$ has the Schwarz function in $B_r(\zeta_0)$ if and only if $\hat{\Omega}$ can be extended holomorphically from $B_r(\zeta_0)\backslash\Omega$ onto $B_r(\zeta_0)$, where $\hat{\Omega}$ denotes the Cauchy transform $\hat{\Omega}(\zeta) = \int_\Omega 1/(z-\zeta)\,dx\,dy$ $(z = x+iy)$ of the characteristic function of Ω.*

Proof. We write B for $B_r(\zeta_0)$. Assume that $\Omega \cup \Gamma$ has the Schwarz function S in B, and let $\tilde{S}$ be the function defined by $\tilde{S}(z) = S(z)$ in Ω and $\tilde{S}(z) = \bar{z}$ on $B\backslash\Omega$. Set $f(z) = \hat{\Omega}(z) + \pi\bar{z} - \pi\tilde{S}(z)$. Then f is continuous in B and holomorphic in $B\backslash\Gamma = \Omega \cup (B\backslash(\Omega \cup \Gamma))$. Since Γ consists of real analytic arcs or is contained in the union of real analytic arcs by our Regularity Theorem, f is holomorphic in B and it is the holomorphic extension of $\hat{\Omega}$ from $B\backslash\Omega$ onto B.

Conversely, if a holomorphic function f in B satisfies $f(z)=\hat{\Omega}(z)$ on $B\backslash\Omega$, then $\hat{\Omega}(z)+\pi\bar{z}-f(z)$ is holomorphic in Ω and continuous on $\Omega\cup\Gamma$. Since $\Gamma\subset B\backslash\Omega$, $(\hat{\Omega}(z)+\pi\bar{z}-f(z))/\pi=\bar{z}$ on Γ. Hence $(\hat{\Omega}(z)+\pi\bar{z}-f(z))/\pi$ is the Schwarz function of $\Omega\cup\Gamma$ in B. □

Remark. From our Regularity Theorem we see that $\tilde{S}$ is Lipschitz continuous on $\overline{B_\delta(\zeta_0)}$ for every δ less than r; see [11, Corollary 5.5]. The proposition, together with the above fact, asserts that if $\hat{\Omega}$ can be extended holomorphically from $B_r(\zeta_0)\backslash\Omega$ onto $B_r(\zeta_0)$, then $\hat{\Omega}$ itself is Lipschitz continuous on $\overline{B_\delta(\zeta_0)}$ for every δ less than r.

We shall prove the following proposition.

Proposition 1.6. *Let Ω be a bounded quadrature domain of a complex measure μ such that* $\operatorname{supp}\mu$ *is contained in Ω. Then there is a global Schwarz function of Ω holomorphic in $\Omega\backslash\operatorname{supp}\mu$. Conversely, if there is a global Schwarz function of a nonempty bounded open set Ω holomorphic in $\Omega\backslash K$, then, for every neighborhood U of K with $\bar{U}\subset\Omega$, there is a complex measure μ such that* $\operatorname{supp}\mu$ *is contained in $\bar{U}$, and Ω is a quadrature domain of μ.*

Proof. If Ω is a bounded quadrature domain of μ such that $\operatorname{supp}\mu\subset\Omega$, then $(\hat{\Omega}(z)+\pi\bar{z}-\hat{\mu}(z))/\pi$ is a global Schwarz function of Ω holomorphic in $\Omega\backslash\operatorname{supp}\mu$. Conversely, if there is a global Schwarz function S of Ω holomorphic in $\Omega\backslash K$, then, from the proof of Proposition 1.5, we see that $f(z)=\hat{\Omega}(z)+\pi\bar{z}-\pi\tilde{S}(z)$, where $\tilde{S}(z)=S(z)$ in $\Omega\backslash K$ and $\tilde{S}(z)=\bar{z}$ on $\mathbb{C}\backslash\Omega$, is holomorphic in $\mathbb{C}\backslash K$. For a neighborhood U of K with $\bar{U}\subset\Omega$, take a modified function g of f such that g is of class C^1 in $\mathbb{C}$ and satisfies $g(z)=f(z)$ in $\mathbb{C}\backslash\bar{U}$. Set $\mu=-(1/\pi)(\partial g/\partial\bar{z})m$, where m denotes the two-dimensional Lebesgue measure. Then $\partial\hat{\mu}/\partial\bar{z}=\partial g/\partial\bar{z}$ in $\mathbb{C}$ and $\operatorname{supp}\mu$ is contained in $\bar{U}$. Hence $\hat{\mu}-g$ is holomorphic in $\mathbb{C}$ and tends to 0 as $z\to\infty$. Therefore, $\hat{\mu}=g$ in $\mathbb{C}$, and so

$$\hat{\mu}(z)=f(z)=\hat{\Omega}(z) \tag{1.1}$$

on $\mathbb{C}\backslash\Omega$. Since $\operatorname{supp}\mu\subset\bar{U}\subset\Omega$, $|\mu|(\mathbb{C}\backslash\Omega)=0$ and $\int|h|d|\mu|<+\infty$ for every holomorphic and integrable function h in Ω. Equation (1.1) implies that

$$\int\frac{1}{z-\zeta}\,d\mu(z)=\int_\Omega\frac{1}{z-\zeta}\,dm(z)$$

for every $\zeta\in\mathbb{C}\backslash\Omega$. Since every holomorphic and integrable function in Ω can be approximated in the mean by linear combinations of $1/(z-\zeta_j)$ with $\zeta_j\in\partial\Omega$; see Bers [2],

$$\int h\,d\mu=\int_\Omega h\,dm$$

holds for every holomorphic and integrable function h in Ω. Thus Ω is a quadrature domain of μ. □

Now we shall show regularity of the boundary of a bounded quadrature domain Ω of a complex measure μ such that $\operatorname{supp}\mu$ is contained in Ω.

Theorem 1.7. *Let Ω be a bounded quadrature domain of a complex measure μ such that* $\operatorname{supp}\mu\subset\Omega$. *Then*

(1) *Every boundary point of Ω is a regular, degenerate, double, or cusp point in the sense of the Regularity Theorem;*

(2) *Every nonisolated degenerate boundary point determines uniquely a regular real analytic simple arc, which tends to* $\operatorname{supp}\mu$ *and does not tend to the set of nondegenerate boundary points of Ω, or determines uniquely a regular real analytic simple closed curve,*

which does not meet supp μ *and the set of nondegenerate boundary points of* Ω. *The number of such arcs and closed curves is finite*;

(3) *There is at most a finite number of isolated degenerate boundary points that do not belong to the union of arcs and closed curves mentioned in* (2);

(4) *The union* $[\Omega]$ *of* Ω *and the set of degenerate boundary points is also a quadrature domain of* μ, *and* $\partial[\Omega]$ *consists of a finite number of real analytic closed curves having at most a finite number of double and cusp points.*

Proof. Let ζ_0 be a nonisolated degenerate boundary point of a bounded quadrature domain Ω. Then, by (2) of the Regularity Theorem, there is a regular real analytic simple arc J passing through ζ_0. Let $\zeta \in J$ and let Ψ be a holomorphic and univalent function defined in $B_\varepsilon(0)$ such that $J \subset \Psi((-\varepsilon, \varepsilon))$ and $\Psi(0) = \zeta$. Then $S_\Psi = \bar{\Psi} \circ \Psi^{-1}$, where $\bar{\Psi}(w) = \overline{\Psi(\bar{w})}$, is the Schwarz function of $(B_\delta(\zeta) \backslash J) \cup (J \cap B_\delta(\zeta))$ at ζ for some $\delta > 0$, and J can be extended onto $\Psi((-\varepsilon, \varepsilon))$.

We note that S_Ψ is equal to the global Schwarz function on Ω in a neighborhood of ζ. We extend J as much as possible. We denote it again by J. Then J has no endpoints in $\Omega \backslash \operatorname{supp} \mu$ by the above argument. J does not tend to the set of nondegenerate boundary points by the Regularity Theorem. Thus J is a regular real analytic simple arc that tends to supp μ or is a closed curve that does not meet supp μ.

Next assume that there is an infinite number of such arcs or closed curves J_j. On each J_j, we can find a nonisolated boundary points ζ_j. By choosing a subsequence if necessary, we may assume that ζ_j converges to a point $\zeta_0 \in \partial\Omega$ because $\partial\Omega$ is compact. By the Regularity Theorem, ζ_0 should be a nonisolated degenerate boundary point. Hence there exists j_0 such that $J_j = J_{j_0}$ for $j \geqq j_0$. This is a contradiction, and there is at most a finite number of such arcs and closed curves.

Finally assume that there is an infinite number of isolated degenerate boundary points ζ_j that do not belong to the union of such arcs and closed curves. Again we may assume that ζ_j converges to a boundary point ζ_0. By the Regularity Theorem, ζ_0 is a nonisolated degenerate boundary point and, for large j, ζ_j must be contained in the same arc or closed curve passing through ζ_0. This is again a contradiction. Hence the number of such ζ_j is finite. □

To show the converse of Theorem 1.7, we introduce and discuss the cluster set of a regular real analytic simple arc. Let J be a regular real analytic simple arc, namely, let ψ be a holomorphic and univalent function defined in a neighborhood of the open interval $(-1, 1)$, and let $J = \psi((-1, 1))$. We denote by C_J the union of $\cap \{\overline{\psi((-1, -1+\delta))};\ \delta > 0\}$ and $\cap \{\overline{\psi((1-\delta, 1))};\ \delta > 0\}$ and call it the cluster set of J. If J is a regular real analytic simple closed curve, then we interpret C_J as the empty set. It follows that $\bar{J} \backslash J \subset C_J \subset \bar{J}$.

Let J be a regular real analytic simple arc as in Theorem 1.7. Then, by our Regularity Theorem, we see that $J \cap C_J = \varnothing$. Namely, we obtain $C_J = \bar{J} \backslash J$. By the same reason, we see that if $J_1, \ldots, J_n$ are regular real analytic simple arcs or closed curves as in Theorem 1.7, then $J_j \cap \overline{J_k} = \varnothing$ for every distinct j and k.

THEOREM 1.8. *Let* $[\Omega]$ *be a bounded open set surrounded by a finite number of real analytic closed curves with at most a finite number of double and cusp points in the sense of our Regularity Theorem. Let* $J_1, \ldots, J_n$ *be regular real analytic simple arcs or closed curves such that* $\overline{J_j} \subset [\Omega]$ *and* $C_{J_j} = \overline{J_j} \backslash J_j$ *for every* j, *and* $J_j \cap \overline{J_k} = \varnothing$ *for every distinct* j *and* k. *Let* Ω *be an open subset of* $[\Omega]$ *such that* $[\Omega] \backslash \Omega$ *is a compact subset of* $\cup J_j$. *Then* Ω *is a quadrature domain of a complex measure* μ *such that* supp μ *is contained in* Ω.

Proof. At each boundary point ζ of Ω, by our Regularity Theorem, we can find a disk $B_\rho(\zeta)$ such that there is a Schwarz function of $(\Omega \cap B_\rho(\zeta)) \cup ((\partial\Omega) \cap B_\rho(\zeta))$ at

ζ. By continuing them, we obtain a global Schwarz function of Ω holomorphic in $\Omega\backslash K$, where K is a compact subset of Ω. Take a neighborhood U of K such that $\bar{U}\subset\Omega$. Then, by Proposition 1.6, Ω is a quadrature domain of a complex measure μ such that $\operatorname{supp}\mu\subset\bar{U}\subset\Omega$. □

Let Ω be a bounded quadrature domain of μ such that supp μ is contained in Ω, and let $[\Omega]$ be the union of Ω and the set of degenerate boundary points of Ω. Then every open set Ω' satisfying $\Omega\subset\Omega'\subset[\Omega]$ is a quadrature domain of μ. Hence Theorems 1.7 and 1.8 give a complete condition for a bounded open set Ω to be a quadrature domain of a complex measure μ such that supp μ is contained in Ω.

II. Unbounded quadrature domains.

2. A Schwarz function of an unbounded set. In this section, we shall discuss a Schwarz function S of an unbounded set in the complex plane $\mathbb{C}$ and show, in Corollary 2.6, that $1/S(1/z)$ is a Schwarz function at 0.

First we define a Schwarz function of an unbounded set. We denote by B_δ a disk with radius δ and center 0.

DEFINITION 2.1. Let Ω be an unbounded open subset of $\mathbb{C}\backslash\overline{B_\delta}$ such that $\Gamma=(\partial\Omega)\backslash\overline{B_\delta}$ is not empty. A function S defined on $\Omega\cup\Gamma$ is called a Schwarz function of $\Omega\cup\Gamma$, more precisely, a Schwarz function of $\Omega\cup\Gamma$ at ∞ or in $\mathbb{C}\backslash\overline{B_\delta}$ if

(i) S is holomorphic in Ω;

(ii) S is continuous on $\Omega\cup\Gamma$;

(iii) $S(\zeta)=\bar{\zeta}$ on Γ;

(iv) $|S(z)|\leqq\alpha|z|$ in $\Omega\backslash\overline{B_{\delta_0}}$ for some $\alpha>0$ and some $\delta_0>\delta$.

We note that condition (iv) is equivalent to the following if Γ is unbounded.

(iv′) $|S(z)|\leqq\alpha|z|^\beta$ in $\Omega\backslash\overline{B_{\delta_0}}$ for some constants α, $\beta>0$, and some $\delta_0>\delta$. Indeed, we apply the Fuchs theorem, Theorem 2.1′ in [11], to $S(z)/z$ and obtain $\lim_{z\in\Omega},\sup_{z\to\infty}|S(z)|/|z|\leqq 1$.

Next we note that the Schwarz function S is determined uniquely if Γ is an infinite set. If there is an accumulation point of Γ in $\mathbb{C}\backslash\overline{B_\delta}$, then the uniqueness follows immediately; see the note given after the definition of a Schwarz function for a bounded set, Definition 3.1, in [11]. In the case that there are no accumulation points of Γ in $\mathbb{C}\backslash\overline{B_\delta}$, we apply Corollary 2.6 below. Then $1/S(1/z)$ is a Schwarz function at 0, and so S is determined uniquely.

In the definition, we do not assume that $S(z)\to\infty$ as $z\in\Omega\to\infty$. But it is true if Γ is unbounded.

PROPOSITION 2.2. *Let S be the Schwarz function of $\Omega\cup\Gamma$ in $\mathbb{C}\backslash\overline{B_\delta}$. If Γ is unbounded, then*

$$|S(z)|>\frac{|z|}{5}\quad in\ \Omega\backslash\overline{B_R}$$

for a large disk B_R.

Remark. If Γ is bounded, then the conclusion of the lemma does not hold, in general. For example, $S(z)=1/z$ is the Schwarz function of $(\mathbb{C}\backslash\overline{B_1})\cup\partial B_1$ in $\mathbb{C}\backslash\overline{B_\delta}$ with $0<\delta<1$ and $\lim_{z\to\infty}S(z)=0$.

To prove Proposition 2.2, we introduce a function $F_c(z)=z(S(z)-cz)$ for a fixed complex number c with $|c|<1$. We set

$$m_c(r)=\inf\{|F_c(z)|;\ z\in\Omega\cap\partial B_r\}$$

for $r>\delta$.

LEMMA 2.3. *It follows that*

(1) *F_c is not constant in any connected component of $\Omega\backslash\overline{B_{\delta_1}}$ for some $\delta_1>\delta$;*

(2) *The integral $\int_{\Omega\cap\partial B_r} d\arg F_c$ is well defined and it is finite if $F_c(z)\neq 0$ on $\Omega\cap\partial B_r$.*

Proof. To prove (1), assume that F_c is identically equal to a constant k in a connected component D of Ω. Then $F_c(\zeta)=|\zeta|^2-c\zeta^2=|\zeta|^2(1-c\,e^{i2\theta})=k$ on $(\partial D)\backslash\overline{B_\delta}$, where $\zeta=|\zeta|\,e^{i\theta}$. Namely, $|\zeta|^2 w=k$, where $w=1-c\,e^{i2\theta}$. Since $(\partial D)\backslash\overline{B_\delta}$ is not empty and $w\neq 0$, $k\neq 0$. If $c\neq 0$, then the line $\{w;\ \arg w=\arg k\}$ and the circle $\{w;\ |w-1|=|c|\}$ crosses at most in two points. Hence $e^{2i\theta}$ has at most two solutions to the above equation, and so $\zeta=|\zeta|\,e^{i\theta}$ has at most four solutions. Thus $(\partial D)\backslash\overline{B_\delta}$ consists of at most four points and $D=\Omega$. This contradicts that Γ is unbounded. Hence $c=0$ and $|\zeta|^2=k$. This means that $k>0$ and $(\partial D)\backslash\overline{B_\delta}\subset\partial B_{\sqrt{k}}$. Since Γ is unbounded, $(\partial D)\backslash\overline{B_\delta}=\partial B_{\sqrt{k}}$ and D is an annulus surrounded by ∂B_δ and $\partial B_{\sqrt{k}}$. We note that such an annulus as well as k exists at most one. We obtain the required δ_1 by taking δ_1 so that $\delta_1\geqq\sqrt{k}$.

Next we prove (2). Since $F_c(z)\to F_c(\zeta)=|\zeta|^2(1-c\,e^{i2\theta})$ as $z\in\Omega\to\zeta=|\zeta|\,e^{i\theta}\in\partial\Omega$, $m_c(r)>0$ if $F_c(z)\neq 0$ on $\Omega\cap\partial B_r$. Let ds denote the line element of ∂B_r. Then

$$d\arg F_c=\frac{\partial\arg F_c}{\partial s}\,ds=\frac{\partial\log|F_c|}{\partial r}\,ds$$

along $\Omega\cap\partial B_r$ and

$$\left|\frac{\partial\log|F_c|}{\partial r}\right|\leqq|(\log F_c)'|=\frac{|F_c'|}{|F_c|}.$$

It follows that $F_c'(z)=(zS(z)-cz^2)'=zS'(z)+S(z)-2cz$. Corollary 5.4 in [11] asserts that $\lim_{z\in\Omega,z\to\zeta}S'(z)$ exists for every ζ on Γ and it is equal to $S'(\zeta)$ if ζ is an isolated point of Γ, and its modulus is equal to one if ζ is not an isolated point of Γ. Hence S' is bounded on $\Omega\cap\partial B_r$. Since $|F_c(z)|\geqq m_c(r)>0$, we see that $(\partial\log|F_c|)/(\partial r)$ is bounded on $\Omega\cap\partial B_r$. Thus the integral $\int_{\Omega\cap\partial B_r}|d\arg F_c|$ is finite. This proves (2). □

Let δ_0 be the number as in (iv) of Definition 2.1, and let δ_1 be the number as in (1) of Lemma 2.3. Take ρ so that $\rho>\max\{\delta_0,\delta_1\}$ and $F_0(z)=zS(z)\neq 0$ on $\Omega\cap\partial B_\rho$. We note that $m_0(\rho)>0$ and $\int_{\Omega\cap\partial B_\rho}d\arg F_0$ is finite. Take ε so that

$$0<\varepsilon<\tfrac{1}{5},\tag{2.1}$$

$$m=\inf\{m_c(\rho);\ |c|\leqq\varepsilon\}>0,\tag{2.2}$$

$$\iota=\sup\left\{\left|\int_{\Omega\cap\partial B_\rho}d\arg F_c\right|;\ |c|\leqq\varepsilon\right\}<+\infty.\tag{2.3}$$

We fix ρ and ε.

Now take λ so that $0<\lambda<\min\{m,\rho^2(1-3\varepsilon)\}$, take c so that $|c|\leqq\varepsilon$, and set

$$\Lambda_{\lambda,c}=\{z\in\Omega\backslash\overline{B_\rho};\ |F_c(z)|=\lambda\}.$$

Since $|F_c(z)|\geqq m>\lambda$ on $\Omega\cap\partial B_\rho$ by (2.2), $\Omega\cap\partial B_\rho\cap\overline{\Lambda_{\lambda,c}}=\varnothing$. Let

$$V=\cup\ \{B_{3\varepsilon t^2}(t^2);\ t\geqq\rho\},$$

take a neighborhood $U(\zeta)$ of $\zeta\in\Gamma\backslash B_\rho$ so that $U(\zeta)\subset B_{(1/3)|\zeta|}(\zeta)$ and $F_0(\Omega\cap U(\zeta))\subset B_{\varepsilon|\zeta|^2}(|\zeta|^2)$, and set

$$U=\cup\ \{U(\zeta);\ \zeta\in\Gamma\backslash B_\rho\}.$$

Since $|F_c(z)-F_0(z)|=|cz^2|\leqq\varepsilon(1+1/3)^2|\zeta|^2<2\varepsilon|\zeta|^2$ in $\Omega\cap U(\zeta)$, $F_c(\Omega\cap U)\subset V$.

Take $r>\rho$, and let $\{\Omega_n\}$ be a regular exhaustion of Ω such that there is at most one connected component of $\Omega_n\cap\partial B_\rho$ (respectively, $\Omega_n\cap\partial B_r$) on each connected component of $\Omega\cap\partial B_\rho$ (respectively, $\Omega\cap\partial B_r$). Take Ω_n so that $(\Omega\backslash\Omega_n)\cap(\overline{B_r}\backslash B_\rho)\subset U$. Since $F_c(\Omega\cap U)\subset V$ and $\operatorname{dist}(V,0)=\rho^2(1-3\varepsilon)>\lambda$, it follows that $(\Omega\backslash\Omega_n)\cap(\overline{B_r}\backslash B_\rho)\cap\overline{\Lambda_{\lambda,c}}=\varnothing$. Let

$$\Omega_{\lambda,c}(r,n)=\{z\in\Omega_n\cap(B_r\backslash\overline{B_\rho});\ |F_c(z)|>\lambda\}.$$

We may assume that $\partial\Omega_{\lambda,c}(r, n)$ consists of a finite number of piecewise real analytic simple closed curves. We note that the number of components of $\partial\Omega_{\lambda,c}(r, n)$, which are entirely contained in $\Lambda_{\lambda,c}$, does not depend on the choice of n if we take Ω_n so that $(\Omega\backslash\Omega_n)\cap(\overline{B_r}\backslash B_\rho)\subset U$. We denote it by $\nu_{\lambda,c}(r)$. Set

$$\Omega_{\lambda,c}=\{z\in\Omega\backslash\overline{B_\rho};\ |F_c(z)|>\lambda\}.$$

Lemma 2.4. *It follows that*

$$\nu_{\lambda,c}(r)\leqq\frac{1}{2\pi}\int_{\Omega_{\lambda,c}\cap\partial B_r} d\arg F_c+\gamma, \tag{2.4}$$

where $\gamma=(\iota+24\operatorname{Arcsin}\varepsilon)/(2\pi)$ *and* ι *is the number defined by* (2.3). *The number* γ *does not depend on the choice of* λ, c, *and* r.

Proof. We apply the argument principle to F_c in $\Omega_{\lambda,c}(r, n)$:

$$\int_{\partial\Omega_{\lambda,c}(r,n)} d\arg F_c=0.$$

It follows that $\int_J d\arg F_c\leqq 0$ for every arc J contained in $\Lambda_{\lambda,c}\cap\partial\Omega_{\lambda,c}(r, n)$. If a component J of $\partial\Omega_{\lambda,c}(r, n)$ is entirely contained in $\Lambda_{\lambda,c}$, then $\int_J d\arg F_c$ is equal to a positive integer multiple of -2π. Hence we obtain

$$\begin{aligned}0&=\int_{\partial\Omega_{\lambda,c}(r,n)} d\arg F_c\\ &\leqq -2\pi\nu_{\lambda,c}(r)-\int_{(\partial B_\rho)\cap\partial\Omega_{\lambda,c}(r,\,n)} d\arg F_c+\int_{(\partial B_r)\cap\partial\Omega_{\lambda,c}(r,n)} d\arg F_c\\ &\quad+\int_{(\partial\Omega_n)\cap\partial\Omega_{\lambda,c}(r,n)} d\arg \mathrm{F}_{\mathrm{c}}.\end{aligned} \tag{2.5}$$

Next we note that $F_c(\partial\Omega_n\cap(\overline{B_r}\backslash B_\rho))\subset V$. This implies that if J is a closed curve contained entirely in $(\partial\Omega_n)\cap\partial\Omega_{\lambda,c}(r, n)$, then $\int_J d\arg F_c=0$. For arcs in $(\partial\Omega_n)\cap\partial\Omega_{\lambda,c}(r, n)$, we may assume that all endpoints of arcs in $(\partial\Omega_n)\cap\partial\Omega_{\lambda,c}(r, n)$ are contained in $(\partial B_\rho)\cup(\partial B_r)$. We write ∂_n for the union of the arcs in $(\partial\Omega_n)\cap\partial\Omega_{\lambda,c}(r, n)$. We note that $|F_c(z)|>\lambda$ on $\Omega\cap\partial B_\rho$. Let $(\Omega\cap\partial B_\rho)_n$ (respectively, $(\Omega_{\lambda,c}\cap\partial B_r)_n$) be the union of connected components of $\Omega\cap\partial B_\rho$ (respectively, $\Omega_{\lambda,c}\cap\partial B_r$), which contains a connected component of $\Omega_n\cap\partial B_\rho$ (respectively, $(\partial\Omega_r)\cap\partial\Omega_{\lambda,c}(r, n)$), and set $\partial_n'=\partial_n+(\Omega\cap\partial B_\rho)_n-\Omega_n\cap\partial B_\rho-(\Omega_{\lambda,c}\cap\partial B_r)_n+(\partial B_r)\cap\partial\Omega_{\lambda,c}(r, n)$. Then, from (2.5), we obtain

$$2\pi\nu_{\lambda,c}(r)\leqq-\int_{(\Omega\cap\partial B_\rho)_n} d\arg F_c+\int_{(\Omega_{\lambda,c}\cap\partial B_r)_n} d\arg F_c+\int_{\partial_n'} d\arg F_c.$$

To estimate the integral on ∂_n', we consider three types of arcs contained in ∂_n: (1) an arc whose one endpoint is on ∂B_ρ and the other is on ∂B_r, (2) an arc whose two endpoints are on ∂B_ρ, and (3) an arc whose two endpoints are on ∂B_r. Let J be an arc of type (1), and let $\rho\, e^{i\varphi}$ and $r\, e^{i\psi}$ be the initial point and the terminal point of J, respectively. Then, by the assumption that there is at most one component of $\Omega_n\cap\partial B_\rho$ (respectively, $\Omega_n\cap\partial B_r$) on each connected component of $\Omega\cap\partial B_\rho$ (respectively, $\Omega\cap\partial B_r$), we can find uniquely determined circular arcs $K(\rho)\subset(\Omega\cap\partial B_\rho)_n$ and $K(r)\subset(\Omega_{\lambda,c}\cap\partial B_r)_n$ such that the initial points of $K(\rho)$ and $K(r)$ are on $\partial\Omega$ and the terminal points of $K(\rho)$ and $K(r)$ are equal to $\rho\, e^{i\varphi}$ and $r\, e^{i\psi}$, respectively. Namely, we can assign an arc $J'=K(\rho)+J-K(r)$ to J and regard it as an arc contained in ∂_n'. We apply the same argument to arcs of type (1) whose initial points are on ∂B_r, arcs of type (2) and arcs of type (3).

We shall estimate the integrals on arcs of type (1). Let $J_1, \ldots, J_{2l}$ be arcs of type (1), and let $\rho\, e^{i\varphi_j}$ and $r\, e^{i\psi_j}$ be endpoints of J_j, $j = 1, 2, \ldots, 2l$. We may assume that $0 \leqq \varphi_1 < \varphi_2 < \cdots < \varphi_{2l} < 2\pi$, $\psi_1 < \psi_2 < \cdots < \psi_{2l} < \psi_1 + 2\pi$, $\rho\, e^{i\varphi_j}$ is the initial point of J_j for odd j and $\rho\, e^{i\varphi_j}$ is the terminal point of J_j for even j. If $\rho\, e^{i\varphi_j}$ is the initial point of J_j, we can find $\varphi_j' < \varphi_j$ such that the circular arc $K_j(\rho)$ between $\rho\, e^{i\varphi_j'}$ and $\rho\, e^{i\varphi_j}$ are in $(\Omega \cap \partial B_\rho)_n$ and $\rho\, e^{i\varphi_j'} \in \partial\Omega$. We can also find $\psi_j' < \psi_j$ such that the circular arc $K_j(r)$ between $r\, e^{i\psi_j'}$ and $r\, e^{i\psi_j}$ are in $(\Omega_{\lambda,c} \cap \partial B_r)_n$ and $r\, e^{i\psi_j'} \in \partial\Omega$. Thus we assign an arc $J_j' = K_j(\rho) + J_j - K_j(r)$ to each J_j for odd j. Similarly, we assign an arc $J_j' = -K_j(r) + J_j + K_j(\rho)$ to each J_j for even j. We note that $F_c(J_j') \subset V$, $F_c(\rho\, e^{i\varphi'}) = \rho^2(1 - c\, e^{2i\varphi'})$ if $\rho\, e^{i\varphi'} \in \partial\Omega$ and $F_c(r\, e^{i\psi'}) = r^2(1 - c\, e^{2i\psi'})$ if $r\, e^{i\psi'} \in \partial\Omega$. We denote Arg w the principal value of arg w, namely, $-\pi < \operatorname{Arg} w \leqq \pi$. Then

$$\int_{J_j'} d \arg F_c = \operatorname{Arg}(1 - c\, e^{2i\psi_j'}) - \operatorname{Arg}(1 - c\, e^{2i\varphi_j'})$$

for odd j and

$$\int_{J_j'} d \arg F_c = \operatorname{Arg}(1 - c\, e^{2i\varphi_j'}) - \operatorname{Arg}(1 - c\, e^{2i\psi_j'})$$

for even j. Hence

$$\sum_{j=1}^{2l} \int_{J_j'} d \arg F_c = \sum_{j=1}^{l} \{\operatorname{Arg}(1 - c\, e^{2i\varphi_{2j}'}) - \operatorname{Arg}(1 - c\, e^{2i\varphi_{2j-1}'})\}$$
$$- \sum_{j=1}^{l} \{\operatorname{Arg}(1 - c\, e^{2i\psi_{2j}'}) - \operatorname{Arg}(1 - c\, e^{2i\psi_{2j-1}'})\}.$$

Since $\varphi_1' < \varphi_2' \leqq \cdots < \varphi_{2l}' \leqq \varphi_1' + 2\pi$ and $\psi_1' < \psi_2' \leqq \cdots < \psi_{2l}' \leqq \psi_1' + 2\pi$, the absolute value of each of the sums in the right-hand side is not greater than 4 Arcsin ε, and so

$$\sum_{j=1}^{2l} \int_{J_j'} d \arg F_c \leqq 8 \operatorname{Arcsin} \varepsilon.$$

Next we shall estimate the integrals on arcs of type (2). For the sake of simplicity, we consider arcs of type (2) contained in a simply connected open set D_1, which is surrounded by J_1, $L_1(r)$, J_2, and $-L_1(\rho)$, where J_1 and J_2 are arcs of type (1), $L_1(r)$ denotes the circular arc between $r\, e^{i\psi_1}$ and $r\, e^{i\psi_2}$, and $L_1(\rho)$ denotes the circular arc between $\rho\, e^{i\varphi_1}$ and $\rho\, e^{i\varphi_2}$.

The situation is complicated because $D_1 \cap \Omega_{\lambda,c}(r, n)$ may not be connected. Let D_1' be a connected component of $D_1 \cap \Omega_{\lambda,c}(r, n)$ such that $\partial D_1' \supset J_1$, and let $I_1, \ldots, I_k$ be arcs of type (2) contained in $\partial D_1'$. Let $\rho\, e^{i\theta_{2j}}$ and $\rho\, e^{i\theta_{2j-1}}$ be the initial point and the terminal point of I_j, $j = 1, \ldots, k$, respectively. We may assume that $\varphi_1 < \theta_1 < \theta_2 < \cdots < \theta_{2k} < \varphi_2$. By using an argument similar to the above, to each I_j, we assign an arc I_j' having endpoints $\rho\, e^{i\theta_{2j}'}$ and $\rho\, e^{i\theta_{2j-1}'}$. It follows that $\theta_1 < \theta_1' \leqq \theta_2' < \theta_2 < \cdots < \theta_{2k-1} < \theta_{2k-1}' \leqq \theta_{2k}' < \theta_{2k}$. Hence

$$\sum_{j=1}^{k} \int_{I_j'} d \arg F_c \leqq \sum_{j=1}^{k} \{\operatorname{Arg}(1 - c\, e^{2i\theta_{2j-1}'}) - \operatorname{Arg}(1 - c\, e^{2i\theta_{2j}'})\}. \tag{2.6}$$

Next we consider the case that there are arcs of type (2) contained in a simply connected open set D_2 surrounded by I_1 and the circular arc between $\rho\, e^{i\theta_1}$ and $\rho\, e^{i\theta_2}$. We take a connected component D_2' of $D_1 \backslash \overline{\Omega_{\lambda,c}(r, n)}$ such that $\partial D_2' \supset I_1$, and let $I_1^{(2)}, \ldots, I_{k(2)}^{(2)}$ be arcs of type (2) contained in $(\partial D_2') \backslash I_1$. As before, we can find arcs $I_j^{(2)'}$ and obtain an estimation similar to (2.6). We note here that the direction of $I_j^{(2)'}$

is opposite from that of I_1'. Thus if we add the integrals on all arcs I' assigned to I of type (2) contained in $\overline{D_2}$, then it is not greater than

$$\int_{\theta_1}^{\theta_2} |d \operatorname{Arg}(1-c\,e^{2i\theta})|.$$

Hence

$$\sum_{(2)} \int_{I_j'} d \arg F_c \leqq \int_{\theta=0}^{2\pi} |d \operatorname{Arg}(1-c\,e^{2i\theta})| = 8 \operatorname{Arcsin} \varepsilon,$$

where $\sum_{(2)}$ denotes the sum for arcs of type (2).

The same estimation holds for arcs of type (3), and we obtain

$$2\pi\nu_{\lambda,c}(r) \leqq -\int_{(\Omega\cap\partial B_\rho)_n} d \arg F_c + \int_{(\Omega_{\lambda,c}\cap\partial B_r)_n} d \arg F_c + 24 \operatorname{Arcsin} \varepsilon.$$

By letting $n\to\infty$, we obtain

$$2\pi\nu_{\lambda,c}(r) \leqq \int_{\Omega_{\lambda,c}\cap\partial B_r} d \arg F_c + 2\pi\gamma. \qquad \square$$

Now we apply the argument as in the proof of Theorem 1 in a paper of Fuchs [4] and prove the following.

LEMMA 2.5. *Inequality* (2.4) *implies that*

$$\nu_{\lambda,c}(r) \leqq \gamma + 2$$

for every $r \geqq \rho$.

Proof. Since $d \arg F_c = (\partial \log |F_c(r\,e^{i\theta})|)/(\partial r) r\,d\theta$ along $\Omega_{\lambda,c}\cap\partial B_r$, dividing by r and integrating the inequality at the end of the proof of Lemma 2.4 from ρ to R we get

$$(2.7) \qquad 2\pi \int_\rho^R \frac{\nu_{\lambda,c}(r)}{r}\,dr \leqq 2\pi\gamma \log(R/\rho) + \iint_{re^{i\theta}\in\Omega_{\lambda,c}(R)} \frac{\partial \log |F_c(r\,e^{i\theta})|}{\partial r}\,d\theta\,dr,$$

where $\Omega_{\lambda,c}(R) = \Omega_{\lambda,c} \cap B_R$. We have already seen that S' is bounded in $\Omega\cap(B_R\backslash\overline{B_\rho})$, and so we see that $(\partial \log |F_c|)/(\partial r)$ is bounded in $\Omega_{\lambda,c}(R)$. Hence we can reverse the order of integration with respect to θ and r. We fix θ and consider

$$\int_{re^{i\theta}\in\Omega_{\lambda,c}(R)} \frac{\partial \log |F_c(r\,e^{i\theta})|}{\partial r}\,dr.$$

The set $\Omega_{\lambda,c}(R)\cap\{z; \arg z=\theta\}$ consists of at most a countable number of open segments. The integral on each segment is equal to $\log|F_c(r_2\,e^{i\theta})| - \log|F_c(r_1\,e^{i\theta})|$, where $r_1\,e^{i\theta}$ and $r_2\,e^{i\theta}$ with $r_2 > r_1$ are the endpoints of the segment. The endpoints are contained in $\partial\Omega$, $\Lambda_{\lambda,c}$, or $(\partial B_\rho)\cup(\partial B_R)$.

First we consider the case that $\{z; \arg z=\theta\}$ does not meet $\Lambda_{\lambda,c}$. If the endpoint $r\,e^{i\theta}$ is contained in $\partial\Omega$, then $|F_c(r\,e^{i\theta})| = r^2|1-c\,e^{2i\theta}|$. Hence the integral is not greater than $\log|F_c(\mathrm{R}e^{i\theta})| - \log|F_c(\rho\,e^{i\theta})|$ if $\rho\,e^{i\theta}$ and $\mathrm{R}e^{i\theta}\in\overline{\Omega_{\lambda,c}(R)}$, $2\log R + \log|1-c\,e^{2i\theta}| - \log|F_c(\rho\,e^{i\theta})|$ if $\rho\,e^{i\theta}\in\overline{\Omega_{\lambda,c}(R)}$ and $\mathrm{R}e^{i\theta}\notin\overline{\Omega_{\lambda,c}(R)}$, $\log|F_c(\mathrm{R}e^{i\theta})| - 2\log\rho - \log|1-c\,e^{2i\theta}|$ if $\rho\,e^{i\theta}\notin\overline{\Omega_{\lambda,c}(R)}$ and $\mathrm{R}e^{i\theta}\in\overline{\Omega_{\lambda,c}(R)}$, and $2\log R - 2\log\rho$ if $\rho\,e^{i\theta}$ and $\mathrm{R}e^{i\theta}\notin\overline{\Omega_{\lambda,c}(R)}$. Since $|F_c(\rho\,e^{i\theta})|\geqq m$ by (2.2), the integral is bounded from above by $\log|F_c(\mathrm{R}e^{i\theta})| + \gamma_1$ or $2\log R + \gamma_1$, where γ_1 denotes a constant independent of the choice of λ, c, R, and θ.

Next we consider the case that $\{z; \arg z=\theta\}$ meets $\Lambda_{\lambda,c}$. Since $\Lambda_{\lambda,c}\cap B_R$ consists of a finite number of real analytic arcs and closed curves, $\{z; \arg z=\theta\}$ meets $\Lambda_{\lambda,c}$ finite times. If the endpoint $r\,e^{i\theta}$ is contained in $\Lambda_{\lambda,c}$, then $|F_c(r\,e^{i\theta})|=\lambda$. We note that $\Omega\cap\partial B_\rho\cap\overline{\Lambda_{\lambda,c}}=\varnothing$ and see that the integral is not greater than $\log|F_c(\mathrm{Re}^{i\theta})|+\gamma_1+\log^+\lambda\leqq\log|F_c(\mathrm{Re}^{i\theta})|+\gamma_1+\log^+ m$ or $2\log R+\gamma_1+\log^+ m$, where $\log^+\lambda=\max\{\log\lambda, 0\}$.

Since $|S(z)\gamma\leqq\alpha|z|$ in $\Omega\backslash\overline{B_{\delta_0}}$ for some α by (iv) of Definition 2.1, we obtain $|F_c(z)|=|z|\,|S(z)-cz|\leqq(\alpha+\varepsilon)|z|^2$ in $\Omega\backslash\overline{B_{\delta_0}}$. Setting $\gamma_2=\gamma_1+\log^+(\alpha+\varepsilon)+\log^+ m$, we obtain

$$\int_{re^{i\theta}\in\Omega_{\lambda,c}(R)}\frac{\partial\log|F_c(r\,e^{i\theta})|}{\partial r}\,dr\leqq 2\log R+\gamma_2.$$

By (2.7),

$$\int_\rho^R\frac{\nu_{\lambda,c}(r)}{r}\,dr\leqq(\gamma+2)\log R+\gamma_3,$$

where $\gamma_3=\gamma_2-\gamma\log\rho$.

Since $\nu_{\lambda,c}(r)$ is a nondecreasing function of r, this inequality is possible for large R only if $\nu_{\lambda,c}(r)\leqq\gamma+2$ for every $r\geqq\rho$. □

Thus we can find R such that all closed curves of $\Lambda_{\lambda,c}$ are contained in $\Omega\cap B_R$. In the above argument, $\gamma+2$ does not depend on the choice of λ and c, but R may depend on λ and c. The final step of our argument is to show that we can take R so that R is independent of the choice of λ and c.

Proof of Proposition 2.2. Set

$$\nu=\sup\{\nu_{\lambda,c}(r);\ r\geqq\rho,\ 0<\lambda<\min\{m,\rho^2(1-3\varepsilon)\},\ |c|\leqq\varepsilon\}$$

and take r_0, λ_0, and c_0 so that $\nu_{\lambda_0,c_0}(r_0)=\nu$. By using the same argument as in the proof of Lemma 3.3 in [11], for fixed r_0 and λ_0, we can find $\varepsilon_1>0$ such that $\nu_{\lambda_0,c}(r_0)=\nu$ for every c in $B_{\varepsilon_1}(c_0)\cap\overline{B_\varepsilon}$. Take $\varepsilon_2>0$ and c_1 so that $B_{\varepsilon_2}(c_1)\subset B_{\varepsilon_1}(c_0)\cap\overline{B_\varepsilon}$. By the definition of r_0, we see that $F_c(z)\neq 0$ in $\Omega\backslash\overline{B_{r_0}}$ for every c in $B_{\varepsilon_2}(c_1)$, namely, $S(z)\neq cz$ in $\Omega\backslash\overline{B_{r_0}}$ for every c in $B_{\varepsilon_2}(c_1)$. This implies that $|S(z)-c_1z|\geqq\varepsilon_2|z|$ in $\Omega\backslash\overline{B_{r_0}}$. We note that $\Gamma=(\partial\Omega)\backslash\overline{B_\delta}$ is unbounded, and we apply the Fuchs theorem, Theorem 2.1′ in [11], to $z/(S(z)-c_1z)$ in $\Omega\backslash\overline{B_{r_0}}$. We see that there exists $R>r_0$ such that $|z/(S(z)-c_1z)|\leqq 2/(1-\varepsilon)$ in $\Omega\backslash\overline{B_R}$. By (2.1), we obtain

$$|S(z)|>\frac{|z|}{5}\quad\text{in }\Omega\backslash\overline{B_R}.$$ □

COROLLARY 2.6. *Let S be the Schwarz function of $\Omega\cup\Gamma$ in $\mathbb{C}\backslash\overline{B_\delta}$. If Γ is unbounded, then*

$$S_i(z)=\begin{cases}\dfrac{1}{\overline{S(1/\bar z)}} & z\in(1/(\Omega\cup\Gamma))\cap B_r,\\ 0 & z=0,\end{cases}$$

is the Schwarz function of $((1/\Omega)\cap B_r)\cup(((1/\Gamma)\cap B_r)\cup\{0\})$ at 0 for some $r<1/\delta$, where $1/E$ for a set E denotes the inversion $\{1/z;\ z\in E\backslash\{0\}\}$ of E.

Proof. Take R as in Proposition 2.2. Then $|S(z)|>|z|/5$ in $\Omega\backslash\overline{B_R}$, and so $1/S(1/z)$ is holomorphic in $1/(\Omega\backslash\overline{B_R})$, is continuous on $1/((\Omega\cup\Gamma)\backslash\overline{B_R})$, and tends to 0 as $z\in 1/(\Omega\backslash\overline{B_R})\to 0$. On $1/\Gamma$, $1/S(1/\zeta)=1/(\overline{1/\zeta})=\bar\zeta$. Hence, S_i is the Schwarz function of $((1/\Omega)\cap B_r)\cup(((1/\Gamma)\cap B_r)\cup\{0\})$ at 0, where $r\leqq 1/R<1/\delta$. □

3. Regularity of boundaries of unbounded quadrature domains. In this section, we discuss regularity of boundaries of unbounded quadrature domains. First we shall define the generalized Cauchy transform of an unbounded open set. For two distinct and fixed points, ζ_1 and ζ_2 in $\mathbb{C}$, we define the generalized Cauchy kernel $K(z, \zeta; \zeta_1, \zeta_2)$ by

$$K(z,\zeta;\zeta_1,\zeta_2)=\frac{(\zeta-\zeta_1)(\zeta-\zeta_2)}{(z-\zeta)(z-\zeta_1)(z-\zeta_2)}$$
$$=\frac{1}{z-\zeta}+\frac{\zeta_2-\zeta}{\zeta_1-\zeta_2}\frac{1}{z-\zeta_1}+\frac{\zeta_1-\zeta}{\zeta_2-\zeta_1}\frac{1}{z-\zeta_2}$$

if $\zeta\in\mathbb{C}\backslash\{\zeta_1,\zeta_2\}$, and

$$K(z,\zeta;\zeta_1,\zeta_2)=0$$

if $\zeta=\zeta_1$ or ζ_2. We set

$$\hat{\mu}(\zeta;\zeta_1,\zeta_2)=\int K(z,\zeta;\zeta_1,\zeta_2)\,d\mu(z)$$

for a complex measure μ and call it the generalized Cauchy transform of μ. It is finite almost everywhere in $\mathbb{C}$ if

$$\iint_{\operatorname{supp}\mu\times\operatorname{supp}\mu}|K(z,\zeta;\zeta_1,\zeta_2)|d|\mu|(z)\,dm(\zeta)<+\infty,$$

where m denotes the two-dimensional Lebesgue measure, and it is holomorphic in the complement of supp μ because the total measure of a complex measure is finite. In contrast with the Cauchy transform, the generalized Cauchy transform can be also defined for a bounded measurable function in $\mathbb{C}$ that does not have compact support. For a bounded measurable function g in $\mathbb{C}$, we define the generalized Cauchy transform $\hat{\hat{g}}$ by

$$\hat{\hat{g}}(\zeta;\zeta_1,\zeta_2)=\int K(z,\zeta;\zeta_1,\zeta_2)g(z)\,dm(z).$$

We write $\hat{\hat{g}}(z)$ for $\hat{\hat{g}}(z;\zeta_1,\zeta_2)$. It is a continuous function in $\mathbb{C}$, vanishes at ζ_1 and ζ_2, and satisfies

$$\frac{\partial\hat{\hat{g}}}{\partial\bar{z}}=-\pi g$$

in the sense of distributions. Furthermore, it satisfies

$$|\hat{\hat{g}}(z)|\leqq\alpha|z|\log|z|$$

for large $|z|$, where α is a positive constant; see, e.g., Chapter IV of Kra [7]. We write $\hat{\hat{\Omega}}$ for $\hat{\hat{\chi}}_\Omega$, where Ω is an open set in $\mathbb{C}$, and χ_Ω denotes the characteristic function of Ω. It is continuous in $\mathbb{C}$ and holomorphic in $\mathbb{C}\backslash\bar{\Omega}$. The function $\hat{\hat{\Omega}}(z)+\pi\bar{z}$ is holomorphic in Ω.

Definition 3.1. Let Ω be an unbounded open proper subset of $\mathbb{C}\backslash\overline{B_\delta}$. Let g be a function continuous on $(\mathbb{C}\backslash\overline{B_\delta})\backslash\Omega$ and holomorphic in the interior of $(\mathbb{C}\backslash\overline{B_\delta})\backslash\Omega$. We say that g can be extended holomorphically from $(\mathbb{C}\backslash\overline{B_\delta})\backslash\Omega$ onto $\mathbb{C}\backslash\overline{B_\delta}$ with at most a simple pole at ∞ if there is a holomorphic function f in $\mathbb{C}\backslash\overline{B_\delta}$ such that

(i) $f(z)=g(z)$ on $(\mathbb{C}\backslash\overline{B_\delta})\backslash\Omega$;

(ii) $|f(z)|\leqq\alpha|z|$ in $\mathbb{C}\backslash\overline{B_{\delta_0}}$ for some $\alpha>0$ and some $\delta_0>\delta$.

Proposition 3.2. *Let Ω be an unbounded open subset of $\mathbb{C}\backslash\overline{B_\delta}$ such that $\Gamma = (\partial\Omega)\backslash\overline{B_\delta}$ is not empty. Then $\Omega\cup\Gamma$ has the Schwarz function at ∞ if and only if $\hat{\Omega}$ can be extended holomorphically from $(\mathbb{C}\backslash\overline{B_\delta})\backslash\Omega$ onto $\mathbb{C}\backslash\overline{B_\delta}$ with at most a simple pole at ∞.*

Proof. Let S be the Schwarz function of $\Omega\cup\Gamma$ at ∞, and let $\tilde{S}$ be the function defined by $\tilde{S}(z) = S(z)$ in Ω and $\tilde{S}(z) = \bar{z}$ on $(\mathbb{C}\backslash\overline{B_\delta})\backslash\Omega$. Set $f(z) = \hat{\Omega}(z) + \pi\bar{z} - \pi\tilde{S}(z)$. Then, by the same argument as in the proof of Proposition 1.3, we see that f is holomorphic in $\mathbb{C}\backslash\overline{B_\delta}$. By definition, $f = \hat{\Omega}$ on $(\mathbb{C}\backslash\overline{B_\delta})\backslash\Omega$. Since $|\hat{\Omega}(z)| \leqq \alpha_1|z|\log|z|$ and $|S(z)| \leqq \alpha_2|z|$ for large $|z|$ by (iv) of Definition 2.1, we see that $|f(z)| \leqq \alpha_3|z|\log|z|$ for large $|z|$, where α_1, α_2, and α_3 are positive constants. Hence f has at most a simple pole at ∞ and satisfies (ii) of Definition 3.1.

Conversely, if a holomorphic function f in $\mathbb{C}\backslash\overline{B_\delta}$ satisfies (i) for $g = \hat{\Omega}$ and (ii) of Definition 3.1, then $S(z) = (\hat{\Omega}(z) + \pi\bar{z} - f(z))/\pi$ is holomorphic in Ω, is continuous on $\Omega\cup\Gamma$, is equal to $\bar{z}$ on Γ, and satisfies

$$|S(z)| \leqq \frac{\alpha_1|z|\log|z| + (\pi+\alpha)|z|}{\pi} \leqq \alpha_2|z|\log|z|,$$

for large $|z|$, where α, α_1, and α_2 are positive constants. If Γ is bounded, then S has at most a simple pole at ∞. Hence $|S(z)| \leqq \alpha_3|z|$ for large $|z|$. If Γ is unbounded, then, by the note mentioned after Definition 2.1, we also obtain $|S(z)| \leqq \alpha_4|z|$ for large $|z|$. Here α_3 and α_4 are positive constants. Hence S is the Schwarz function of $\Omega\cup\Gamma$ at ∞. □

Holomorphic extensions of the Cauchy transforms of open sets are closely related to quadrature domains. We shall prove the following two lemmas.

Lemma 3.3. *Let Ω be a quadrature domain of a complex measure μ. Let $\zeta_0 \in (\partial\Omega)\backslash\operatorname{supp}\mu$, and let $B = B_r(\zeta_0)$ be a disk such that $B\cap\operatorname{supp}\mu = \varnothing$. Then $(\Omega\cap B)^\wedge$ can be extended holomorphically from $B\backslash\Omega$ onto B.*

Proof. We note that Ω may be unbounded. We may assume that $B\backslash\Omega$ contains more than two points. Take two distinct points, ζ_1 and ζ_2, on $B\backslash\Omega$, write $\hat{\mu}(\zeta)$ for $\hat{\mu}(\zeta; \zeta_1, \zeta_2)$, and write $\hat{E}$ for $\hat{\chi}_E$, where E denotes a Borel set in $\mathbb{C}$. Since the generalized Cauchy kernel $K(z, \zeta; \zeta_1, \zeta_2)$ given before Definition 3.1 is holomorphic and integrable on Ω for each fixed $\zeta \in \mathbb{C}\backslash\Omega$, we obtain $\hat{\mu}(\zeta) = \hat{\Omega}(\zeta)$ on $B\backslash\Omega$. Since $\hat{\Omega} = (\Omega\cap B)^\wedge + (\Omega\backslash B)^\wedge$,

$$(\Omega\cap B)^\wedge = \hat{\Omega} - (\Omega\backslash B)^\wedge = \hat{\mu} - (\Omega\backslash B)^\wedge$$

on $B\backslash\Omega$. By the assumption that $B\cap\operatorname{supp}\mu = \varnothing$, the right-hand side is holomorphic in B, and so $(\Omega\cap B)^\wedge$ can be extended holomorphically from $B\backslash\Omega$ onto B. We obtain the lemma because $(\Omega\cap B)^\wedge(\zeta) = (\Omega\cap B)^\wedge(\zeta) + a\zeta + b$, where $a = ((\Omega\cap B)^\wedge(\zeta_1) - (\Omega\cap B)^\wedge(\zeta_2))/(\zeta_2 - \zeta_1)$ and $b = (\zeta_1(\Omega\cap B)^\wedge(\zeta_2) - \zeta_2(\Omega\cap B)^\wedge(\zeta_1))/(\zeta_2 - \zeta_1)$ are constants. □

Lemma 3.4. *Let Ω be a quadrature domain of a complex measure μ with compact support. Let $B = B_R$ be a disk satisfying $\operatorname{supp}\mu \subset \bar{B}$, and assume that $(\mathbb{C}\backslash\bar{B})\backslash\Omega$ contains two distinct points, ζ_1 and ζ_2. Set $(\Omega\backslash\bar{B})^\wedge(\zeta) = (\Omega\backslash\bar{B})^\wedge(\zeta; \zeta_1, \zeta_2)$. Then $(\Omega\backslash\bar{B})^\wedge$ can be extended holomorphically from $(\mathbb{C}\backslash\bar{B})\backslash\Omega$ onto $\mathbb{C}\backslash\bar{B}$ with at most a simple pole at ∞.*

Proof. Since $K(z, \zeta; \zeta_1, \zeta_2)$ is holomorphic and integrable on Ω for every fixed ζ on $\mathbb{C}\backslash\Omega$, we obtain

$$(\Omega\backslash\bar{B})^\wedge(\zeta) = \hat{\mu}(\zeta; \zeta_1, \zeta_2) - (\Omega\cap\bar{B})^\wedge(\zeta; \zeta_1, \zeta_2)$$

on $(\mathbb{C}\backslash\bar{B})\backslash\Omega$. The right-hand side of the equality is a holomorphic function of ζ in

$\mathbb{C}\setminus\bar{B}$ because supp $\mu\subset\bar{B}$ and neither ζ_1 nor ζ_2 is contained in $\bar{B}$. Since $|K(z,\zeta;\zeta_1,\zeta_2)|\leqq\alpha|\zeta|$ for large $|\zeta|$ and for z on $\bar{B}$, we obtain $|\hat{\hat{\mu}}(\zeta;\zeta_1,\zeta_2)|\leqq\alpha_1|\zeta|$ and $|(\Omega\cap\bar{B})\hat{\ }(\zeta;\zeta_1,\zeta_2)|\leqq\alpha_2|\zeta|$ for large $|\zeta|$, where α, α_1, and α_2 are positive constants. Thus, by definition, $(\Omega\setminus\bar{B})\hat{\ }$ can be extended holomorphically from $(\mathbb{C}\setminus\bar{B})\setminus\Omega$ onto $\mathbb{C}\setminus\bar{B}$ with at most a simple pole at ∞. □

Now we apply Corollary 2.6 and show regularity of the boundary of an unbounded quadrature domain.

PROPOSITION 3.5. *Let Ω be an unbounded quadrature domain of a complex measure μ. Then, for every $\zeta\in(\partial\Omega)\setminus$*supp *$\mu$, there is a disk $B=B_\delta(\zeta)$ and one of the statements* (1)–(4) *of the Regularity Theorem holds. Furthermore, under the assumption that the support of μ is compact, we can find a disk $B=B_\delta$ such that one of the statements* (1)–(4) *of the Regularity Theorem holds, if we replace Ω and Γ by $(1/\Omega)\cap B_{1/\rho}$ and $((1/\Gamma)\cap B_{1/\rho})\cup\{0\}$, respectively, where ρ is chosen so that* supp $\mu\subset\overline{B_\rho}$.

Proof. The first assertion follows from Lemma 3.3, Proposition 1.5, and the Regularity Theorem. For the second assertion it is trivial if the boundary of Ω is bounded. If $\partial\Omega$ is unbounded, then, by Lemma 3.4 and Proposition 3.2, $(\Omega\setminus\overline{B_\rho})\cup(\Gamma\setminus\overline{B_\rho})$ has the Schwarz function S at ∞. By Corollary 2.6, $S_i(z)$ is the Schwarz function of $((1/\Omega)\cap B_{1/\rho})\cup(((1/\Gamma)\cap B_{1/\rho})\cup\{0\})$ at 0. The proposition follows from our Regularity Theorem. □

COROLLARY 3.6. *Let Ω be an unbounded quadrature domain of a complex measure with compact support. Then* (area $(\Omega\cap B_r))/(\pi r^2)$ *converges to $\frac{1}{2}$ or 1 as $r\to+\infty$.*

The corollary was conjectured by Shapiro [12] and proved under some regularity hypotheses on $\partial\Omega$. From the corollary, we can easily obtain the following result: If a quadrature domain Ω of a complex measure with compact support has infinite area, then $\int_{\Omega\setminus B_1}|z|^{-2}\,dm(z)=+\infty$; see in [9, Thm. 11.2].

Now we discuss an unbounded quadrature domain Ω of a complex measure μ with compact support such that $(\partial\Omega)\cap\operatorname{supp}\mu=\varnothing$; in other words, a quadrature domain Ω of μ such that supp μ is compact and contained in Ω. We introduce a global Schwarz function of an unbounded open set.

DEFINITION 3.7. Let Ω be an unbounded open set in $\mathbb{C}$. A function S defined on $\bar{\Omega}\setminus K$ for some compact subset K of Ω is called a global Schwarz function of Ω, or a global Schwarz function of Ω holomorphic in $\Omega\setminus K$, if it satisfies (i)–(iii) of Definition 1.3 and

(iv) $|S(z)|\leqq\alpha|z|$ in $\Omega\setminus B_R$ for some positive numbers α and R.

The global Schwarz function of Ω holomorphic in $\Omega\setminus K$ is determined uniquely, if $\Omega\setminus K$ is connected and if $\partial\Omega$ is an infinite set, see the note after Definition 3.1 in [11].

Next we shall prove an "unbounded" version of Proposition 1.6.

PROPOSITION 3.8. *Let Ω be an unbounded quadrature domain of a complex measure μ such that* supp μ *is compact and contained in Ω. Then there is a global Schwarz function of Ω holomorphic in $\Omega\setminus$*supp *μ. Conversely, if there is a global Schwarz function of an unbounded open set Ω holomorphic in $\Omega\setminus K$, then, for every relatively compact neighborhood U of K with $\bar{U}\subset\Omega$, there is a complex measure μ such that* supp μ *is contained in $\bar{U}$, and Ω is a quadrature domain of μ.*

Proof. If $\mathbb{C}\setminus\Omega$ consists of at most two points, then the family of holomorphic and integrable functions on Ω consists of just one function, the constant function with value zero. See [9, Lemma 11.3]. Hence Ω is a quadrature domain of any measure μ such that $|\mu|(\mathbb{C}\setminus\Omega)=0$. In this case, a global Schwarz function of Ω always exists and is not determined uniquely. For example, every rational function $r(z)$ having at most a simple pole at ∞ and satisfying $r(\zeta)=\bar{\zeta}$ on $\mathbb{C}\setminus\Omega$ is a global Schwarz function on Ω. Hence the proposition holds in this case.

Thus we may assume that $\mathbb{C}\backslash\Omega$ consists of more than two points. Take two distinct points ζ_1 and ζ_2 on $\mathbb{C}\backslash\Omega$, write $\hat{\hat{\mu}}(\zeta)$ for $\hat{\hat{\mu}}(\zeta;\zeta_1,\zeta_2)$ and write $\hat{\hat{E}}$ for $\hat{\hat{\chi}}_E$. If Ω is a quadrature domain of μ such that $\operatorname{supp}\mu$ is compact and contained in Ω, then, $S(z)=(\hat{\hat{\Omega}}(z)+\pi\bar{z}-\hat{\hat{\mu}}(z))/\pi$ satisfies (i)-(iii) of Definition 1.3 for $K=\operatorname{supp}\mu$. By the argument in the proof of Proposition 3.2, we see that S also satisfies (iv) of Definition 3.7, and so S is a global Schwarz function of Ω holomorphic in $\Omega\backslash\operatorname{supp}\mu$.

Conversely, assume that there exists a global Schwarz function S of Ω holomorphic in $\Omega\backslash K$. By Propositions 1.5 and 3.2, $f(z)=\hat{\hat{\Omega}}(z)+\pi\bar{z}-\pi\tilde{S}(z)$ is holomorphic in $\mathbb{C}\backslash K$, has at most a simple pole at ∞, and vanishes at ζ_1 and ζ_2, where $\tilde{S}$ denotes the function defined by $\tilde{S}(z)=S(z)$ in Ω and $\tilde{S}(z)=\bar{z}$ on $\mathbb{C}\backslash\Omega$. For a relatively compact neighborhood U of K with $\bar{U}\subset\Omega$, take a modified C^1-function g of f as in the proof of Proposition 1.6. Then we see that $\mu=-(1/\pi)(\partial g/\partial\bar{z})m$ satisfies $\partial\hat{\hat{\mu}}/\partial\bar{z}=\partial g/\partial\bar{z}$ in $\mathbb{C}$ and $\operatorname{supp}\mu\subset\bar{U}$. Since $\hat{\hat{\mu}}-g$ is a linear function in $\mathbb{C}$ and vanishes at ζ_1 and ζ_2, $\hat{\hat{\mu}}=g$ in $\mathbb{C}$. Hence $\hat{\hat{\mu}}(z)=f(z)=\hat{\hat{\Omega}}(z)$ on $\mathbb{C}\backslash\Omega$. We apply again the Bers approximation theorem replaced the Cauchy kernel $1/(z-\zeta)$ with the generalized Cauchy kernel $K(z,\zeta;\zeta_1,\zeta_2)$ and see that Ω is a quadrature domain of μ satisfying $\operatorname{supp}\mu\subset\bar{U}$. □

Now we shall show regularity of the boundary of an unbounded quadrature domain Ω of a complex measure μ such that $\operatorname{supp}\mu$ is compact and contained in Ω.

THEOREM 3.9. *Let Ω be an unbounded quadrature domain of a complex measure μ such that* $\operatorname{supp}\mu$ *is compact and contained in Ω. Then, by considering Ω and $[\Omega]$ in the topology of the Riemann sphere, we obtain the same consequence as in Theorem* 1.7. *Conversely, let $[\Omega]$ be an open subset of the Riemann sphere surrounded by a finite number of real analytic closed curves with at most a finite number of double and cusp points in the sense of the Regularity Theorem. Let $J_1,\dots,J_n$ be regular real analytic simple arcs or closed curves such that $\bar{J_j}\subset[\Omega]$ and $C_{J_j}=\bar{J_j}\backslash J_j$ for every j, and $J_j\cap\bar{J_k}=\varnothing$ for every distinct j and k. Let Ω be an open subset of $[\Omega]\cap\mathbb{C}$ such that $[\Omega]\backslash\Omega$ is a compact subset of $\bigcup J_j$ in the topology of the Riemann sphere. Then Ω is a quadrature domain of a complex measure μ such that* $\operatorname{supp}\mu$ *is compact in $\mathbb{C}$ and contained in Ω.*

Proof. By using the same argument as in the proof of Theorem 1.7, we obtain the first assertion of the theorem from Proposition 3.8. For the converse, we apply the same argument as in the proof of Theorem 1.8 and Proposition 3.8 if the point at infinity is not an isolated boundary point of Ω in the topology of the Riemann sphere. If ∞ is an isolated boundary point of Ω, then we take a function S so that S is holomorphic outside of a large disk and has at most a simple pole at ∞. At each finite boundary point of Ω, we take a Schwarz function as in the proof of Theorem 1.8. Then S becomes a global Schwarz function of Ω, and Ω is a quadrature domain of a complex measure μ such that $\operatorname{supp}\mu$ is compact and contained in Ω by Proposition 3.8. □

We shall finally prove the following.

THEOREM 3.10. *Let Ω be an unbounded quadrature domain of a complex measure μ such that* $\operatorname{supp}\mu$ *is compact and contained in Ω. Then*

(1) *$\bar{\Omega}=\mathbb{C}$ and all boundary points of Ω are degenerate in the sense of the Regularity Theorem*; *or*

(2) *Ω is a translation of the inversion of a bounded quadrature domain Ω_0 of some complex measure μ_0 such that* $\operatorname{supp}\mu_0\subset\Omega_0$. *Here the inversion of Ω_0 is the image of $\Omega_0\backslash\{0\}$ under a mapping $w\mapsto 1/w$.*

Proof. If $\bar{\Omega}=\mathbb{C}$, then, by Proposition 3.5, all boundary points of Ω are degenerate. Assume that $\mathbb{C}\backslash\bar{\Omega}\neq\varnothing$ and take a point ζ_0 in $\mathbb{C}\backslash\bar{\Omega}$. Set $w=T(z)=1/(z-\zeta_0)$, and let S be a global Schwarz function of Ω holomorphic in $\Omega\backslash K$ of which existence is guaranteed by Proposition 3.8.

If $\partial\Omega$ is unbounded, then, since S is continuous on $\partial\Omega$ and $S(z)\to\infty$ as $z\in\Omega\to\infty$ by Proposition 2.2, we can take K so that $S(z)\neq\overline{\zeta_0}$ in $\Omega\backslash K$. Set $S_T=\bar{T}\circ S\circ T^{-1}$, where $\bar{T}(z)=1/(z-\overline{\zeta_0})$. Then, S_T is a global Schwarz function of a bounded open set $\Omega_0=T(\Omega)$ holomorphic in $\Omega_0\backslash T(K)$. If $\partial\Omega$ is bounded, then we again take K so that $S(z)\neq\overline{\zeta_0}$ in $\Omega\backslash K$. $S_T=\bar{T}\circ S\circ T^{-1}$ is a global Schwarz function of a bounded open set $\Omega_0=T(\Omega)\cup\{0\}$ holomorphic in $\Omega_0\backslash T(K)\backslash\{0\}$.

By Proposition 1.6, we see that Ω_0 is a quadrature domain of some μ_0 such that supp $\mu_0\subset\Omega_0$. Ω is a translation of the inversion $1/(\Omega_0\backslash\{0\})$ of Ω_0 because $z=T^{-1}(w)=1/w+\zeta_0$. □

4. Quadrature domains of point differential functionals of finite order. In this section, we shall discuss "classical" quadrature domains, namely, quadrature domains of representing measures μ of point differential functionals

$$L(f)=\sum_{j=1}^{l}\sum_{k=1}^{n_j}a_{jk}f^{(k-1)}(z_j),$$

where z_j are distinct points and a_{jk} are constants independent of functions f. We assume that $a_{jn_j}\neq 0$ for every j and put $n=\sum_{j=1}^{l}n_j$. The natural number n is called the order of L.

Definition 4.1. Let Ω be an open set in $\mathbb{C}$. We call Ω a quadrature domain of L if

(i) $\{z_1,\ldots,z_l\}\subset\Omega$;

(ii) For every holomorphic and integrable function f in Ω,

$$L(f)=\int_\Omega f\,dm,$$

where m denotes the two-dimensional Lebesgue measure. We say that Ω is a quadrature domain of order n if Ω is a quadrature domain of some functional L of order n.

If Ω is a quadrature domain of L, then we can find a complex measure μ such that supp $\mu\subset\Omega$ and

$$L(f)=\int f\,d\mu$$

for every holomorphic function f in Ω. Hence Ω is a quadrature domain of the complex measure μ.

Definition 4.2. A nonempty open set Ω is called a null quadrature domain or a quadrature domain of order 0 if

$$\int_\Omega f\,dm=0$$

for every holomorphic and integrable function f in Ω.

Null quadrature domains are called quadrature domains of order 0 by Shapiro [12]; see Theorems 4.4 and 4.6 below. All null quadrature domains are unbounded. They are completely determined by the author [8]. We shall call Ω a quadrature domain of finite order if Ω is a quadrature domain of order n for a natural number n or a null quadrature domain. By Propositions 1.6 and 3.8 a quadrature domain Ω of finite order has a global Schwarz function of Ω. We note that it is determined uniquely if $\partial\Omega$ is an infinite set. The following lemma shows that the global Schwarz function of a quadrature domain of finite order can be extended meromorphically onto Ω. In what follows, we call the extended meromorphic function the global Schwarz function of Ω.

LEMMA 4.3. *A nonempty open set Ω in $\mathbb{C}$ is a quadrature domain of order n if and only if there is a global Schwarz function of Ω that can be extended meromorphically onto Ω and the extended meromorphic function has n poles in Ω.*

Proof. First we note that if $\mathbb{C}\backslash\Omega$ consists of at most two points, then the family of holomorphic and integrable functions in Ω consists of just one function, the constant function with value zero; see [9, Lemma 11.3]. Hence, in this case, Ω is a quadrature domain of any functional L if $\{z_1, \ldots, z_l\} \subset \Omega$. We can easily construct a meromorphic function S such that S has n poles in Ω, has at most a simple pole at ∞, and satisfies $S(\zeta) = \bar{\zeta}$ on $\partial\Omega$. For example, if $\mathbb{C}\backslash\Omega = \{\zeta_1, \zeta_2\}$ and if $n = 0$,

$$S(z) = \bar{\zeta}_1 \frac{z - \zeta_2}{\zeta_1 - \zeta_2} + \bar{\zeta}_2 \frac{z - \zeta_1}{\zeta_2 - \zeta_1}$$

is the required function.

Thus we may assume that $\mathbb{C}\backslash\Omega$ contains more than two points. Let ζ_1 and ζ_2 be two distinct points on $\mathbb{C}\backslash\Omega$. Let $\hat{L}(\zeta; \zeta_1, \zeta_2) = L(K(\cdot, \zeta; \zeta_1, \zeta_2))$, where K denotes the generalized Cauchy kernel. If Ω is a quadrature domain of order n, then by the proof of Proposition 3.8, we see that

$$S(z) = \frac{\hat{\Omega}(z; \zeta_1, \zeta_2) + \pi\bar{z} - \hat{L}(z; \zeta_1, \zeta_2)}{\pi}$$

is a global Schwartz function of Ω. The function $\hat{L}(z; \zeta_1, \zeta_2)$ is meromorphic in Ω and has poles in $\{z_1, \ldots, z_l\}$. At z_j it has a pole of order n_j. Hence S is meromorphic and has n poles in Ω.

Conversely, let S be a global Schwarz function of Ω. By the proof of Proposition 3.8, Ω is a quadrature domain of a complex measure μ such that supp μ is compact and contained in Ω, and S can be expressed as

$$S(z) = \frac{\hat{\Omega}(z; \zeta_1, \zeta_2) + \pi\bar{z} - \hat{\mu}(z; \zeta_1, \zeta_2)}{\pi}$$

in $\Omega\backslash\operatorname{supp}\mu$. If S can be extended meromorphically onto Ω and has n poles in Ω, then it is the same for $\hat{\mu}$. We denote the extension of $\hat{\mu}$ by M. It is a rational function that has n poles in Ω, has at most a simple pole at ∞, and vanishes at ζ_1 and ζ_2. Take a point differential functional L of order n so that $\hat{L}(\zeta, \zeta_1, \zeta_2) = (L(K(\cdot, \zeta; \zeta_1, \zeta_2))$ has the same poles as M has as a function of ζ. Then $\hat{L}(\zeta; \zeta_1, \zeta_2) - M(\zeta)$ is a linear function in $\mathbb{C}$ and vanishes at ζ_1 and ζ_2. Hence $\hat{L}(\zeta; \zeta_1, \zeta_2) = M(\zeta)$ in $\mathbb{C}$, and so

$$L(K(\cdot, \zeta; \zeta_1, \zeta_2)) = M(\zeta) = \hat{\mu}(\zeta; \zeta_1, \zeta_2) = \int K(z, \zeta; \zeta_1, \zeta_2)\, d\mu(z)$$

on $\mathbb{C}\backslash\Omega$. We apply again the Bers approximation theorem in [2] and see that μ is a representing measure of L. □

In contrast with a bounded quadrature domain Ω of an arbitrary complex measure μ satisfying supp $\mu \subset \Omega$, the boundary of a bounded quadrature domain of finite order has a special feature. Indeed, Aharonov and Shapiro proved in [1] that the boundary of a bounded connected quadrature domain of finite order is contained in the set of zeros of an irreducible polynomial of x and y. Gustafsson proved in [5] and [6] that the boundary consists of all zeros of the polynomial except for a finite set, namely, there are no nonisolated degenerate boundary points of a bounded quadrature domain of finite order. The core of the proof is to make use of the Schottky double of a regular domain.

There are satisfactory arguments on bounded quadrature domains of finite order. See the references cited above. Here we shall discuss unbounded quadrature domains of finite order.

First we shall discuss the case $\bar{\Omega}=\mathbb{C}$. Davis proved in his book [3] that if the global Schwarz function of a bounded quadrature domain is a rational function, then it is a linear fractional function. We make use of his idea and prove the following theorem.

THEOREM 4.4. *Let Ω be a quadrature domain of a functional L of order n such that $\bar{\Omega}=\mathbb{C}$ and $\mathbb{C}\backslash\Omega$ consists of more than $(n+1)^2$ points. Then the order is equal to* 0 *and*

(1) *$\mathbb{C}\backslash\Omega$ is contained in a line or the order is equal to* 1; *and*

(2) *$\mathbb{C}\backslash\Omega$ is contained in a circle.*

In the latter case, the functional L is of the form $L(f)=a_{11}f(z_1)$, where z_1 denotes the center of the circle, and a_{11} denotes a positive constant that is equal to the area of the inside of the circle.

In particular, if Ω is a quadrature domain of order n such that $\bar{\Omega}=\mathbb{C}$ and if Ω is neither an open set stated in (1) *nor one stated in* (2), *then $n\neq 0$ and $\mathbb{C}\backslash\Omega$ consists of at most $(n+1)^2$ points.*

Conversely, an open set Ω stated in (1) *is a quadrature domain of order* 0, *and an open set Ω stated in* (2) *is a quadrature domain of $L(f)=a_{11}f(z_1)$, where z_1 denotes the center of the circle and a_{11} denotes the area of the inside of the circle. If Ω is an open set such that $\mathbb{C}\backslash\Omega$ consists of $(n+2)$ points, $n\geqq 0$, then there is a functional L of order at most n such that Ω is a quadrature domain of L.*

Proof. Let Ω be a quadrature domain of a functional L of order n such that $\bar{\Omega}=\mathbb{C}$. By Lemma 4.3, a global Schwarz function S of Ω is meromorphic and has n poles in Ω. By (1) of Theorem 3.10 all boundary points of Ω are degenerate, and so S is holomorphic on $\mathbb{C}\backslash\Omega$. Hence S is a rational function having n poles in $\mathbb{C}$ and it has at most a simple pole at ∞. In other words, S is a rational function of order at most $n+1$ such that the denominator is a polynomial of degree n.

First we shall show that if $\mathbb{C}\backslash\Omega$ consists of more than $(n+1)^2$ points, then $n=0$ or 1 and S is a nonconstant linear fractional function such that $\bar{S}=S^{-1}$, where $\bar{S}(z)=\overline{S(\bar{z})}$. The function $\bar{S}$ is also a rational function of order at most $n+1$ such that the denominator is a polynomial of degree n. The composite $\bar{S}\circ S$ of S and $\bar{S}$ is a rational function of order at most $(n+1)^2$ such that the denominator is a polynomial of degree at most $(n+1)^2-1$, and so it has at most $(n+1)^2$ finite fixed points if it is not the identity. Since $(\bar{S}\circ S)(z)=\bar{S}(S(z))=\bar{S}(\bar{z})=\overline{S(z)}=z$ on $\partial\Omega$, $\bar{S}\circ S$ is the identity mapping of the Riemann sphere if $\mathbb{C}\backslash\Omega$ consists of more than $(n+1)^2$ points. Since $(S\circ\bar{S})(z)=z$ on $\{\bar{z};\, z\in\partial\Omega\}$, by the same reason, we see that $S\circ\bar{S}$ is also the identity mapping of the Riemann sphere. Hence S is a nonconstant linear fractional function such that $\bar{S}=S^{-1}$, and $n=0$ or 1.

Next we shall show that if Ω is a quadrature domain of order 0, then $\partial\Omega$ is contained in a line. The assumption $n=0$ implies that S has no poles in $\mathbb{C}$. Hence $S(z)=az+b$ for some constants $a\neq 0$ and b. Since $\bar{S}(z)=\bar{a}z+\bar{b}$ and $S^{-1}(z)=(1/a)z-b/a$, $\bar{a}=1/a$ and $\bar{b}=-b/a$. Hence $|a|=1$ and $b/\sqrt{-a}$ is real for a square root $\sqrt{-a}$ of $-a$. Set $w=\sqrt{-a}\,z$. Then $az+b=\bar{z}$ if and only if $2\operatorname{Re} w=w+\bar{w}=b/\sqrt{-a}$. Hence $\{z\in\mathbb{C};\, S(z)=az+b=\bar{z}\}$ is a line and $\partial\Omega$ is contained in the line.

Next we shall show that if Ω is a quadrature domain of a functional L of order 1, then $\partial\Omega$ is contained in a circle and L is the functional stated in (2). Since S is a linear fractional function having a pole in Ω, S can be expressed as $S(z)=(az+b)/(z+d)$, where a, b, and d are constants satisfying $ad-b\neq 0$. Since $\bar{S}(z)=(\bar{a}z+\bar{b})/(z+\bar{d})$ and $S^{-1}(z)=(-dz+b)/(z-a)$, we obtain $\bar{a}=-d$, $\bar{b}=b$, and $\bar{d}=-a$.

Set $w=z+d$. Then $(az+b)/(z+d)=\bar{z}$ if and only if $|w|^2=(z+d)(\bar{z}+\bar{d})=(z+d)(\bar{z}-a)=b-ad$. Hence $b-ad>0$, and $\partial\Omega$ is contained in a circle with center $-d$ and radius $\sqrt{b-ad}$. Since

$$S(z)=\frac{az+b}{z+d}=\frac{b-ad}{z+d}+a,$$

we see that

$$\hat{\hat{L}}(z;\zeta_1,\zeta_2)=\frac{-\pi(b-ad)}{z+d}+g(z)$$

in a neighborhood of $z=-d$, where g denotes a holomorphic function in the neighborhood. On the other hand, if $L(f)=a_{11}f(z_1)$, then, by the definition of the generalized Cauchy kernel $K(z,\zeta;\zeta_1,\zeta_2)$,

$$\hat{\hat{L}}(z;\zeta_1,\zeta_2)=\frac{-a_{11}}{z-z_1}+h(z),$$

where $h(z)$ is a linear function of z. Hence $a_{11}=\pi(b-ad)$ and $z_1=-d$, namely, a_{11} is the area of the inside of the circle and z_1 is the center of the circle.

Conversely, if $\mathbb{C}\backslash\Omega$ is contained in a line, then Ω is the union of two half planes and a subset of the line. It is easy to show that each half plane is a null quadrature domain; see the remark at the end of [8]. Hence Ω is a quadrature domain of order 0. If $\mathbb{C}\backslash\Omega$ is contained in a circle $\partial B_r(z_1)$, then Ω is a quadrature domain of $L(f)=\pi r^2 f(z_1)$ because $\mathbb{C}\backslash\overline{B_r(z_1)}$ is a null quadrature domain; see the same remark in [8], and $B_r(z_1)$ is a quadrature domain of $L(f)=\pi r^2 f(z_1)$.

The proof will be complete if we show that there is a functional L of order at most n and Ω is a quadrature domain of L if $\mathbb{C}\backslash\Omega$ consists of $(n+2)$ points. We may assume that $n\geqq 1$. The linear space of all holomorphic and integrable functions in Ω is of dimension n; see [9, Lemma 11.3]. Let $(f_1,\ldots,f_n)$ be a basis of the space, and let $W(f_1,\ldots,f_n)$ be the Wronskian of $(f_1,\ldots,f_n)$. Since $f_1,\ldots,f_n$ are linearly independent, we can choose a point z_1 in Ω so that $W(f_1,\ldots,f_n)(z_1)\neq 0$. Then a system of linear equations

$$\sum_{k=1}^{n} a_{1k}f_j^{(k-1)}(z_1)=\int_\Omega f_j\,dm, \qquad j=1,\ldots,n$$

has a unique solution $(a_{11},\ldots,a_{1n})$. Thus Ω is a quadrature domain of $L(f)=\sum_{k=1}^{n} a_{1k}f^{(k-1)}(z_1)$ of order at most n. □

Remark. In the proof of the proposition, we have chosen a point z_1 in Ω so that $W(f_1,\ldots,f_n)(z_1)\neq 0$. There is an infinite number of such points z_1. Hence there is an infinite number of functionals L such that Ω is a quadrature domain of L if one of $\int_\Omega f_j\,dm$ is not equal to zero. Thus, by deleting a point from Ω if necessary, we see that there are an infinite number of global Schwarz functions of Ω that are meromorphic in Ω if $\partial\Omega$ is a finite set. It is also not difficult to show that if $\mathbb{C}\backslash\Omega$ consists of $n+2$ points, then, for every given $z_1,\ldots,z_l$ in Ω and for every given $n_1,\ldots,n_l$ with $\sum n_j=n$, there exists only one global Schwarz function of Ω that has a pole at z_j of order at most n_j, $j=1,\ldots,l$ and is holomorphic in $\Omega\backslash\{z_1,\ldots,z_l\}$.

Next we give an example of quadrature domain Ω of order 1 such that $\bar{\Omega}=\mathbb{C}$, but $\mathbb{C}\backslash\Omega$ is not contained in a circle. By Theorem 4.4, $\mathbb{C}\backslash\Omega$ consists of four points.

Example. Let $f(z,\zeta)=K(z,\zeta;0,-1)$ be the generalized Cauchy kernel, namely,

$$f(z,\zeta)=\frac{1}{z-\zeta}-\frac{1+\zeta}{z}+\frac{\zeta}{z+1}=\frac{\zeta(\zeta+1)}{(z-\zeta)z(z+1)}. \tag{4.1}$$

First we calculate the integral of $f(z, \zeta)$ on $\mathbb{C}\backslash\{0, -1, \zeta\}$. Let $\partial B_r(c)$ be a circle passing through 0, -1, and ζ. Here we set $\zeta = \xi + i\eta$ and assume that $\eta \neq 0$. The integral is equal to the integral of $f(z, \zeta)$ on $B_r(c)$, and so it is equal to $\pi r^2 f(c, \zeta)$. Substituting c for z in (4.1), we obtain

$$f(c, \zeta) = \frac{\bar{c} - \bar{\zeta}}{|c-\zeta|^2} - \frac{(1+\zeta)\bar{c}}{|c|^2} + \frac{\zeta(\bar{c}+1)}{|c+1|^2}.$$

Since all denominators of fractions in the right-hand side are equal to r^2, we obtain

$$\int f(z, \zeta)\, dm = \pi r^2 f(c, \zeta) = \pi(\zeta - \bar{\zeta}) = 2\pi i \eta.$$

Now we shall take ζ so that $\Omega = \mathbb{C}\backslash\{0, -1, i, \zeta\}$ is the desired example. Let $g(z, \zeta) = f(z, \zeta)/f(z, i)$. If we can choose ζ so that $f(z_1, i) \neq 0$ and $g(z_1, \zeta) = \eta$ for some $z_1 \in \Omega$, then Ω is a quadrature domain of $L(f) = a_{11} f(z_1)$, where $a_{11} = 2\pi i/f(z_1, i)$. Indeed, $(f(z, i), f(z, \zeta))$ is a basis of the class of all holomorphic and integrable functions in Ω. By definition, $a_{11} f(z_1, i) = 2\pi i = \int f(z, i)\, dm$. On the other hand, $a_{11} f(z_1, \zeta) = 2\pi i f(z_1, \zeta)/f(z_1, i) = 2\pi i g(z_1, \zeta) = 2\pi i \eta = \int f(z, \zeta)\, dm$.

Finally we shall show that there is ζ such that $f(z_1, i) \neq 0$ and $g(z_1, \zeta) = \eta$ for some $z_1 \notin \{0, -1, i, \zeta\}$. By (4.1),

$$g(z, \zeta) = \frac{\zeta(\zeta+1)(z-i)}{i(i+1)(z-\zeta)}. \tag{4.2}$$

Hence g is a linear fractional function of z for fixed ζ and has a zero at i and a pole at ζ. Since $\lim_{z\to 0} g(z, \zeta) = (\zeta+1)/(i+1)$ and $\lim_{z\to -1} g(z, \zeta) = \zeta/i$, we can find $z_1 \notin \{0, -1, i, \zeta\}$ if $(\zeta+1)/(i+1) \neq \eta$ and $\zeta/i \neq \eta$. These are equivalent to $\eta \neq \xi + 1$ and $\xi \neq 0$, respectively. Namely, if ζ is not on the union of two lines $\{\eta = \zeta + 1\}$ and $\{\xi = 0\}$, then the unique solution z_1 to $g(z_1, \zeta) = \eta$ does not belong to $\{0, -1, i, \zeta\}$. From (4.2) we obtain

$$z_1 = \frac{\eta i(i+1)\zeta - i\zeta(\zeta+1)}{\eta i(i+1) - \zeta(\zeta+1)}.$$

If we take $\eta = 1$, then

$$z_1 = \frac{-\zeta}{i\zeta + (i-1)}.$$

Hence if we vary ζ on a line $\{\eta = 1\}$, then z_1 moves on a circle, and so $f(z_1, i) \neq 0$ for some $\zeta \neq i$.

For a bounded quadrature domain Ω of finite order, we have already mentioned that there are no nonisolated degenerate boundary points. As we have seen in Theorem 1.7, the union $[\Omega]$ of Ω and the set of isolated degenerate boundary points is also a quadrature domain of the same functional. Let S be the global Schwarz function of Ω, then it is also the global Schwarz function of $[\Omega]$, and it is a meromorphic function having n poles in Ω, where n denotes the order of the functional.

DEFINITION 4.5. Let Ω be a bounded quadrature domain of finite order. We call a point z in $[\Omega]$ a singular point of Ω if $S(z) = \bar{z}$. We call it a simple singular point if z is a simple $\bar{z}$-point of S, namely, if $S(z) = \bar{z}$ and $S'(z) \neq 0$. We call a point z in $[\Omega]$ a regular point of Ω if $S(z) \neq \bar{z}$.

Let Ω be a bounded quadrature domain of finite order. We note that singular points of Ω are isolated degenerate boundary points of Ω or are points in Ω. If ζ is a singular point in Ω, then $\Omega\backslash\{\zeta\}$ is also a quadrature domain of the functional. Hence there is at most a finite number of singular points of Ω. We note that every point in Ω is regular or singular and every boundary point of Ω is a regular, singular, double, or cusp point.

Let Ω be an arbitrary open set in $\mathbb{C}$. We say that Ω is the inversion of Ω_0 with respect to a point ω_0 if Ω is the image of $\Omega_0\backslash\{\omega_0\}$ under a mapping $w \mapsto 1/(w-\omega_0)$. The following proposition is essentially due to Shapiro.

THEOREM 4.6 [12, Thm. 3.3]. *Let Ω be an unbounded quadrature domain of order n such that $\bar{\Omega} \neq \mathbb{C}$. Then*

(1) *Ω is a translation of the inversion of a bounded quadrature domain Ω_0 of order $n+1$ with respect to a regular point or a regular boundary point of Ω_0;*
or

(2) *Ω is a translation of the inversion of a bounded quadrature domain Ω_0 of order $n+2$ with respect to a simple singular point of Ω_0, a double point, or a cusp point of $\partial\Omega_0$.*

Conversely, an open set Ω stated in (1) *or* (2) *is an unbounded quadrature domain of order n such that $\bar{\Omega} \neq \mathbb{C}$.*

Proof. Let ζ_0 be a point in $\mathbb{C}\backslash\bar{\Omega}$, and set $w = T(z) = 1/(z-\zeta_0)$. Then, by the proof of Theorem 3.10, Ω is a translation of the inversion of a bounded open set $\Omega_0 = T(\Omega)$ or $T(\Omega) \cup \{0\}$ with respect to the origin because $z = T^{-1}(w) = 1/w + \zeta_0$. $S_T = \bar{T} \circ S \circ T^{-1}$ is the global Schwarz function of Ω_0 and is meromorphic in Ω_0, where S denotes the global Schwarz function of Ω. Hence Ω_0 is a bounded quadrature domain of finite order. We add all isolated boundary points of Ω_0 to Ω_0 and make a quadrature domain $[\Omega_0]$.

To determine the order of $[\Omega_0]$, we note that $[\Omega_0]$ is of order m if and only if S_T has m poles in $[\Omega_0]$. We also note that S_T has m poles in $[\Omega_0]$ if and only if S has m $\overline{\zeta_0}$-points in $[\Omega] = T^{-1}([\Omega_0])$ because $\bar{T}(z) = 1/(z-\overline{\zeta_0})$. Here, if $0 \in [\Omega_0]$, then we consider $[\Omega]$ in the topology of the Riemann sphere and regard ∞ as a point in $[\Omega]$.

We shall apply the argument principle to $S(z)-\overline{\zeta_0}$ and count the number m of $\overline{\zeta_0}$-points of S in $[\Omega]$. First we consider the case that $\partial\Omega$ is bounded, namely, $[\Omega]$ is an open set containing ∞ in the Riemann sphere. Since $S(z)-\overline{\zeta_0} = \bar{z}-\overline{\zeta_0} \neq 0$ on $\partial[\Omega] \subset \partial\Omega$, the argument principle for domains in the Riemann sphere asserts that

$$m-n = \frac{1}{2\pi}\int_{\partial[\Omega]} d \arg\,(S(z)-\overline{\zeta_0})$$

if S is regular at ∞. Since $S(z)-\overline{\zeta_0} = \bar{z}-\overline{\zeta_0}$ on $\partial[\Omega]$ and ζ_0 belongs to the exterior of $[\Omega]$, we see that

$$\frac{1}{2\pi}\int_{\partial[\Omega]} d \arg\,(S(z)-\overline{\zeta_0}) = \frac{1}{2\pi}\int_{\partial[\Omega]} d \arg\,(\bar{z}-\overline{\zeta_0}) = 1.$$

Hence $m = n+1$. We note that S is regular at ∞ if and only if $S_T(0) \neq 0$. Thus the origin is a regular point of Ω_0. If S has a simple pole at ∞, then, by the argument principle, $m-(n+1) = 1$, and so $m = n+2$. In this case, $S_T(0) = 0$ and the origin is a simple singular point of Ω_0.

Next we consider the case that $\partial\Omega$ is unbounded, namely, 0 is a boundary point of $[\Omega_0]$. In this case, we note that $S(z) \to \infty$ as $z \to \infty$. For large R such that $S(z)-\overline{\zeta_0} \neq 0$ on $[\Omega]\backslash B_R$, we apply the argument principle to $S-\overline{\zeta_0}|[\Omega] \cap B_R$. Then

$$m-n = \frac{1}{2\pi}\int_{\partial([\Omega]\cap B_R)} d \arg\,(S(z)-\overline{\zeta_0}).$$

If 0 is a regular boundary point of $[\Omega_0]$, then, by the Regularity Theorem and Corollary 2.6,

$$\frac{1}{2\pi}\int_{[\Omega]\cap\partial B_R} d\arg\,(S(z)-\overline{\zeta_0}) \tag{4.3}$$

tends to $\frac{1}{2}$ as $R\to+\infty$. Since

$$\frac{1}{2\pi}\int_{\bar{B}_R\cap\partial[\Omega]} d\arg\,(S(z)-\overline{\zeta_o})=\frac{1}{2\pi}\int_{\bar{B}_R\cap\partial[\Omega]} d\arg\,(\bar{z}-\overline{\zeta_0}) \tag{4.4}$$

tends to $\frac{1}{2}$ as $R\to+\infty$, we obtain $m-n=1$. If 0 is a double or cusp point of $\partial[\Omega_0]$, then both (4.3) and (4.4) tend to 1 as $R\to+\infty$, and so $m-n=2$. Hence Ω_0 is of order $n+1$ if 0 is a regular boundary point of Ω_0 and is of order $n+2$ if 0 is a double or cusp point of $\partial\Omega_0$.

To show the converse, let Ω_0 be a bounded quadrature domain of order m, and let $\Omega=T^{-1}(\Omega_0\backslash\{\omega_0\})$, where $z=T^{-1}(w)=1/(w-\omega_0)+a$ denotes a translation of the inversion with respect to ω_0 and ω_0 is a fixed point on $\overline{\Omega_0}$ such that it is simple if it is a singular point. The global Schwarz function S_0 of Ω_0 has m poles and $S=\overline{T^{-1}}\circ S_0\circ T$ is the global Schwarz function of Ω because ω_0 is simple if it is a singular point of Ω. Hence Ω is a quadrature domain of finite order. Let n be the number of poles of S in Ω; here we consider Ω as a subset of $\mathbb{C}$. Then it is equal to the number of $\overline{\omega_0}$-points of S_0 in $[\Omega_0]\backslash\{\omega_0\}$.

If $\omega_0\in[\Omega_0]$ and if ∞ is a removable singularity of S, then $S_0(\omega_0)\neq\overline{\omega_0}$, and so

$$n-m=\frac{1}{2\pi}\int_{\partial[\Omega_0]} d\arg\,(S_0(w)-\overline{\omega_0})=\frac{1}{2\pi}\int_{\partial[\Omega_0]} d\arg\,(\bar{w}-\overline{\omega_0})=-1.$$

If $\omega_0\in[\Omega_0]$ and if ∞ is a simple pole of S, then ω_0 is a simple singular point, and so $(n+1)-m=-1$. If $\omega_0\in\partial[\Omega_0]$, for sufficiently small $\delta>0$, we obtain

$$\begin{aligned}n-m=&\frac{1}{2\pi}\int_{(\partial[\Omega_0])\backslash B_\delta(\omega_0)} d\arg\,(S_0(w)-\overline{\omega_0})\\&+\frac{1}{2\pi}\int_{-(\partial B_\delta(\omega_0))\cap[\Omega_0]} d\arg\,(S_0(w)-\overline{\omega_0}).\end{aligned}$$

The first term of the right-hand side of the equality is equal to

$$\frac{1}{2\pi}\int_{(\partial[\Omega_0])\backslash B_\delta(\omega_0)} d\arg\,(\bar{w}-\overline{\omega_0}).$$

Each term of the right-hand side of the equality tends to $-\frac{1}{2}$ as $\delta\to0$ if ω_0 is a regular point of $\partial[\Omega_0]$ and -1 if ω_0 is a double or cusp point of $\partial[\Omega_0]$. Hence $n-m=-1$ or -2. This completes the proof. □

COROLLARY 4.7. *Let Ω be a quadrature domain of finite order. If there is a nonisolated degenerate boundary point of Ω, then Ω is of order* 0 *and of type* (1) *of Theorem* 4.4 *or Ω is of order* 1 *and of type* (2) *of Theorem* 4.4.

COROLLARY 4.8. *Let Ω be a quadrature domain of finite order such that $\partial\Omega$ is neither a line nor a circle. Then each connected component of the exterior of Ω is not a quadrature domain of finite order.*

From the corollary, we see that the inside of an ellipse, which is not a circle, is not a quadrature domain of finite order because the outside of the ellipse is a quadrature domain of order 0; see § 5.

5. Null quadrature domains. By Theorems 4.4 and 4.6, we can construct all unbounded quadrature domains of order n, if we know all bounded quadrature domains of order $n+1$ and $n+2$. In [8], all quadrature domains of order 0 were determined. They are

(1) A domain whose boundary is a proper subset of a line in $\mathbb{C}$;
(2) A half plane;
(3) The exterior of a circle;
(4) The exterior of an ellipse;
(5) The exterior of a parabola;
(6) The complement of a closed parallel strip.

We shall here apply Theorems 4.4 and 4.6 for $n=0$ and construct all of them.

If $\bar{\Omega}=\mathbb{C}$, then by Theorem 4.4 Ω is a domain stated in (1) or in (6), where the closed parallel strip is a line. If $\bar{\Omega}\neq\mathbb{C}$, then by Theorem 4.6 Ω is a translation of the inversion of a bounded quadrature domain of order 1 or 2. A bounded quadrature domain of order 1 is a disk and all points and all boundary points of the disk are regular. The inversion of a disk with respect to a boundary point is a half plane and the inversion with respect to an interior point is the exterior of a circle. Hence they are domains stated (2) and (3), respectively.

For a bounded connected quadrature domain of order 2, it is known that $[\Omega]$ is simply connected; see Aharonov and Shapiro [1] and (2) of § 5 in [10]. By Corollary to Theorem 2.1 in Gustafsson [6], we see that $c+2d+e\leqq 1$, where c, d, and e denote the number of cusp, double, and singular points of Ω, respectively. Namely, there are no double points and there is at most one cusp or singular point in this case. It is also known that the domain is obtained as the image of the unit disk under a rational function R of order two, which is univalent in the unit disk; see Theorem 1 in Aharonov and Shapiro [1] and Chapter 14 of Davis [3].

If 0 is a singular point of the domain, then, as shown in [10, § 5, eqn. (2)], we may assume that the rational function R has the form $R(w)=w/(1+c_1w+c_2w^2)$, where $0<|c_2|<1$. It is also shown there that $S(z)=(1/\overline{c_2})z+O(z^2)$ in a neighborhood of 0, and so 0 is a simple singular point. Now the boundary of the inversion of the domain with respect to 0 is the image of $\{|w|=1\}$ under $w\mapsto(1+c_1w+c_2w^2)/w$. Since $1/w=\bar{w}$ on $\{|w|=1\}$, the boundary can be expressed as $\{\bar{w}+c_2w+c_1;\ |w|=1\}$. Since $0<|c_2|<1$, it is an ellipse and the inversion is a domain stated in (4).

If there is a cusp point on the boundary of a bounded connected quadrature domain of order two, we may assume that the domain is the image of the unit disk under $R(w)=w(1+w/(2+a))/(1-aw)$, where $|a|<1$, and $R(-1)=-1/(2+a)$ is the cusp; [10, § 5, eqn. (2)]. Hence the boundary of the inversion of the domain with respect to the cusp is the image of $\{|w|=1\}$ under

$$w\mapsto\frac{1}{R(w)-R(-1)}=\frac{(2+a)(1-aw)}{(1+w)^2}.$$

Since $1/w=\bar{w}$ on $\{|w|=1\}$, the right-hand side of the equality is equal to

$$\frac{(2+a)(\bar{w}-a)}{(1+w)(\bar{w}+1)}=\frac{(2+a)(\bar{w}-a)}{2(1+\operatorname{Re} w)}.$$

We set $(\bar{w}-a)/(1+\operatorname{Re} w)=x+iy$, $a=\alpha+i\beta$ and $w=u+iv$. Then $x=(u-\alpha)/(u+1)$ and $y=-(v+\beta)/(u+1)$. Hence we then have $u=-(x+\alpha)/(x-1)$ and $v=-(\beta x-(1+\alpha)y-\beta)/(x-1)$. Since $u^2+v^2=1$ on $\{|w|=1\}$, we obtain $(x+\alpha)^2+(\beta x-(1+\alpha)y-\beta)^2=(x-1)^2$ on $\{|w|=1\}$, namely, $(1+\alpha)(-2x+(1-\alpha))=(\beta x-(1+\alpha)y-\beta)^2$ on $\{|w|=1\}$. Since $|a|<1$, it follows that $1+\alpha\neq 0$. Hence the boundary of the inversion is a parabola. This is the case stated in (5).

Finally, we consider a bounded quadrature domain of order 2 that is not connected. Then each connected component is a bounded quadrature domain of order 1, and so it is a disk. If these two disks do not touch each other, then all boundary points are regular. Hence the only possibility is the case that these two disks touch each other at one point. Then the point is a double point and the inversion with respect to the point is the complement of a closed parallel strip with positive width stated as in (6).

REFERENCES

[1] D. Aharonov and H. S. Shapiro, *Domains on which analytic functions satisfy quadrature identities*, J. Anal. Math., 30 (1976), pp. 39-73.

[2] L. Bers, *An approximation theorem*, J. Anal. Math., 14 (1965), pp. 1-4.

[3] P. J. Davis, *The Schwarz Function and Its Applications*, Carus Math. Monographs, No. 17, Math. Assoc. America, Washington, D.C., 1974.

[4] W. H. J. Fuchs, *A Phragmén-Lindelöf theorem conjectured by D. J. Newman*, Trans. Amer. Math. Soc., 267 (1981), pp. 285-293.

[5] B. Gustafsson, *Quadrature identities and the Schottky double*, Acta Appl. Math., 1 (1983), pp. 209-240.

[6] ———, *Singular and special points on quadrature domains from an algebraic geometric point of view*, J. Anal. Math., 51 (1988), pp. 91-117.

[7] I. Kra, *Automorphic Forms and Kleinian Groups*, Benjamin, Reading, MA, 1972.

[8] M. Sakai, *Null quadrature domains*, J. Anal. Math., 40 (1981), pp. 144-154.

[9] ———, *Quadrature Domains*, Lecture Notes in Math. 934, Springer-Verlag, Berlin, 1982.

[10] ———, *An index theorem on singular points and cusps of quadrature domains*, in Holomorphic Functions and Moduli Vol. I, Math. Sci. Res. Inst. Publ., 10, Springer-Verlag, New York, 1988, pp. 119-131.

[11] ———, *Regularity of a boundary having a Schwarz function*, Acta Math., 166 (1991), pp. 263-297.

[12] H. S. Shapiro, *Unbounded quadrature domains*, in Complex Analysis Vol. I, Lecture Notes in Math. 1275, Springer-Verlag, Berlin, 1987, pp. 287-331.

SIAM J. MATH. ANAL.
Vol. 24, No. 2, pp. 365-389, March 1993

MODIFIED FAR FIELD OPERATORS IN INVERSE SCATTERING THEORY*

DAVID COLTON†‡ AND PETER HÄHNER†§

Abstract. The dual space method for solving the inverse scattering problem reformulates the inverse problem as one in constrained optimization with weighted averages of the far field pattern u_∞ as data. Unfortunately there exists a discrete set of values of the wave number such that the infimum of the cost functional associated with the optimization scheme is not zero. In this paper, the authors show how this difficulty can be removed if instead of u_∞, $u_\infty - u_\infty^0$ is considered, where u_∞^0 is the far field pattern of a surface potential.

Key words. inverse scattering, acoustic waves, electromagnetic waves

AMS(MOS) subject classifications. 35J05, 35P25, 35R30

1. Introduction. In [5] and [6] Colton and Monk introduced a new method for solving the inverse scattering problem for time harmonic acoustic waves that has subsequently been referred to as the dual space method or the method of superposition of incident fields [3]. This method has been extended to the case of time harmonic electromagnetic waves by Blöhbaum [1] and Colton and Päivärinta [9]; see also [3]. The main advantage of the dual space method is that in the case of many incident waves the number of unknowns is significantly reduced by means of an averaging process. On the other hand, a disadvantage of the method is that it fails at a discrete set of values of the wave number corresponding to an interior eigenvalue for the obstacle problem or a transmission eigenvalue for the case of an inhomogeneous medium. In the case of an inhomogeneous medium, this disadvantage was overcome in [7] and [4] by replacing the far field pattern u_∞ by $u_\infty - u_\infty^0$, where u_∞^0 is the far field pattern corresponding to a ball with impedance boundary condition. A more flexible approach for the scalar case was given in [8] where u_∞ is again replaced by $u_\infty - u_\infty^0$, where now u_∞^0 is the far field pattern corresponding to an arbitrary modal expansion with an inequality constraint on the Fourier coefficients.

In this paper, we shall continue this study by examining the possibility of choosing u_∞^0 to be the far field pattern corresponding to a surface potential. It has not escaped our attention that in some ways the present approach resembles the modified integral equation method for the direct scattering problem that is usually associated with the names Brakhage, Leis, Panich, and Werner (cf. [2]). However, as pointed out in [8], the resemblance is somewhat superficial since the failure of the standard integral equation method for solving the direct scattering problem is due to the ansatz, whereas the failure of the dual space method at an eigenvalue is due to the fact that for such values of the wave number the set of far field patterns corresponding to arbitrary incident plane waves is not complete.

Before stating more precisely what will be done in this paper, we formulate the direct scattering problems under consideration. In the following, $D \subset \mathbb{R}^3$ is a bounded domain containing the origin with C^2 boundary ∂D such that $D_e := \mathbb{R}^3 \setminus \bar{D}$ is connected. We denote the unit exterior normal to ∂D (or to the boundary of any other domain)

* Received by the editors April 2, 1992; accepted for publication July 9, 1992.

† Department of Mathematical Sciences, University of Delaware, Newark, Delaware 19716.

‡ The research of this author was supported in part by the Air Force Office of Scientific Research.

§ The research of this author was supported by a scholarship of the Deutsche Forschungsgemeinschaft.

by ν. The function n is a given positive function such that $m := 1 - n$ has compact support and $n \in C^{0,\alpha}(\mathbb{R}^3)$ in the scalar case and $n \in C^{1,\alpha}(\mathbb{R}^3)$ in the vector case, where $C^{0,\alpha}$ and $C^{1,\alpha}$ denote the usual Hölder spaces ($C_0^{0,\alpha}$ and $C_0^{1,\alpha}$ will denote Hölder spaces consisting of functions with compact support). The function n is called the refractive index of the inhomogeneous medium. The incident field in the scalar case is always $u^i(x, d) := \exp(i\kappa x \cdot d)$, $x \in \mathbb{R}^3$, where $\kappa > 0$ is the wave number and $d \in \Omega := \{x \in \mathbb{R}^3 : |x| = 1\}$. For the electromagnetic case the incident field is always given by

$$\begin{aligned} E^i(x, d, p) &:= \frac{i}{\kappa} \operatorname{curlcurl} \{p \exp(i\kappa x \cdot d)\}, \\ H^i(x, d, p) &:= \operatorname{curl} \{p \exp(i\kappa x \cdot d)\}, \qquad x \in \mathbb{R}^3, \end{aligned} \tag{1.1}$$

where $p \in \mathbb{R}^3$ is a constant vector denoting the polarization, E^i is the incident electric field, and H^i is the incident magnetic field. The corresponding scattered fields u^s and E^s, H^s are required to satisfy, respectively, the Sommerfeld radiation condition

$$\lim_{r \to \infty} r\left(\frac{\partial u^s}{\partial r} - i\kappa u^s\right) = 0$$

and the Silver–Müller radiation condition

$$\lim_{r \to \infty} (H^s \times x - rE^s) = 0$$

uniformly for all directions $\hat{x} := x/|x|$, where $r = |x|$. Finally, the total fields $u = u^i + u^s$ and $E = E^i + E^s$, $H = H^i + H^s$, are assumed to satisfy the regularity conditions $u \in C^2(D_e) \cap C(\bar{D}_e)$ and E, $H \in C^1(D_e) \cap C(\bar{D}_e)$ in the case of obstacle scattering and $u \in C^2(\mathbb{R}^3)$ and E, $H \in C^1(\mathbb{R}^3)$ in the case of scattering by an inhomogeneous medium.

We can now state the following direct scattering problems, each of which has a unique solution [3].

Direct acoustic obstacle problem. Find $u = u^i + u^s$ such that

$$\Delta u + \kappa^2 u = 0 \quad \text{in } D_e \tag{1.2}$$

and

$$u = 0 \quad \text{on } \partial D. \tag{1.3}$$

Direct acoustic inhomogeneous medium problem. Find $u = u^i + u^s$ such that

$$\Delta u + \kappa^2 n(x) u = 0 \quad \text{in } \mathbb{R}^3. \tag{1.4}$$

Direct electromagnetic obstacle problem. Find $E = E^i + E^s$ and $H = H^i + H^s$ such that

$$\operatorname{curl} E - i\kappa H = 0, \qquad \operatorname{curl} H + i\kappa E = 0 \quad \text{in } D_e \tag{1.5}$$

and

$$\nu \times E = 0 \quad \text{on } \partial D. \tag{1.6}$$

Direct electromagnetic inhomogeneous medium problem. Find $E = E^i + E^s$ and $H = H^i + H^s$ such that

$$\operatorname{curl} E - i\kappa H = 0, \qquad \operatorname{curl} h + i\kappa n(x) E = 0 \quad \text{in } \mathbb{R}^3. \tag{1.7}$$

Note that since n is assumed positive the inhomogeneous medium for acoustic waves is nonabsorbing and for electromagnetic waves is a dielectric.

Of primary concern in this paper is the inverse scattering problem corresponding to each of the above direct scattering problems. To formulate the inverse problems, we first note that for acoustic waves u^s has the asymptotic behavior

$$u^s(x, d) = \frac{\exp(i\kappa|x|)}{|x|} u_\infty(\hat{x}, d) + O\left(\frac{1}{|x|^2}\right)$$

as $|x| \to \infty$, where u_∞ is the *far field pattern* and for electromagnetic waves E^s has the asymptotic behavior

$$E^s(x, d, p) = \frac{\exp(i\kappa|x|)}{|x|} E_\infty(\hat{x}, d, p) + O\left(\frac{1}{|x|^2}\right)$$

as $|x| \to \infty$, where E_∞ is the *electric far field pattern.* For obstacle scattering, the inverse problem is to determine D from u_∞ or E_∞, and for scattering by an inhomogeneous medium the inverse problem is to determine n (or m) from u_∞ or E_∞. The dual space method reformulates each of these inverse problems as a problem of determining the minimum of a functional over a compact or weakly compact set. For our purposes, it is important to recall what form these functionals take.

Let $\tilde{u}$ be a given radiating solution of the Helmholtz equation (1.2) in $\mathbb{R}^3\backslash\{0\}$ with far field pattern $\tilde{u}_\infty$, and let $\tilde{E}$, $\tilde{H}$ be a given radiating solution of the Maxwell equations (1.5) in $\mathbb{R}^3\backslash\{0\}$ with electric far field pattern $\tilde{E}_\infty$. For $g \in L^2(\Omega)$, define the acoustic far field operator $\mathbf{F}_a$ by

$$(\mathbf{F}_a g)(\hat{x}) := \int_\Omega u_\infty(\hat{x}, d) g(d)\, ds(d), \qquad \hat{x} \in \Omega,$$

and the acoustic Herglotz wave function v by

$$v(x) := \int_\Omega \exp(i\kappa x \cdot d) g(d)\, ds(d), \qquad x \in \mathbb{R}^3.$$

For $g \in L^2(\Omega)$ a tangential vector field, define the electromagnetic far field operator $\mathbf{F}_e$ by

$$(\mathbf{F}_e g)(\hat{x}) := \int_\Omega E_\infty(\hat{x}, d, g(d))\, ds(d), \qquad \hat{x} \in \Omega,$$

and the electric Herglotz wave function E^i_g by

$$\begin{aligned} E^i_g(x) &:= \int_\Omega E^i(x, d, g(d))\, ds(d) \\ &= i\kappa \int_\Omega \exp(i\kappa x \cdot d) g(d)\, ds(d), \qquad x \in \mathbb{R}^3. \end{aligned}$$

Then the (unmodified) dual space method for solving the inverse scattering problem can be formulated as follows, where S is the set of C^2 surfaces that bound a domain D containing the origin such that D_e is connected and B is an open ball containing the support of $m := 1 - n$.

Inverse acoustic obstacle problem. Find $g \in L^2(\Omega)$ and $\Gamma \in S$ that minimize the functional

$$M_1(g, \Gamma) := \|\mathbf{F}_a g - \tilde{u}_\infty\|_{L^2(\Omega)} + \|\tilde{u} + v\|_{L^2(\Gamma)}.$$

Inverse acoustic inhomogeneous medium problem. Find $g \in L^2(\Omega)$, $m \in C_0^{0,\alpha}(B)$ and $w \in C^2(\bar{B})$ a solution of (1.4) in B that minimize the functional

$$M_2(g, w, m) := \|\mathbf{F}_a g - \tilde{u}_\infty\|_{L^2(\Omega)} + \|w - \tilde{u} - v\|_{L^2(\partial B)} + \left\| \frac{\partial}{\partial r}(w - \tilde{u} - v) \right\|_{L^2(\partial B)}.$$

Inverse electromagnetic obstacle problem. Find a tangential vector field $g \in L^2(\Omega)$ and $\Gamma \in S$ that minimize the functional

$$M_3(g, \Gamma) := \|\mathbf{F}_e g - \tilde{E}_\infty\|_{L^2(\Omega)} + \|\nu \times (\tilde{E} + E_g^i)\|_{L^2(\Gamma)}.$$

Inverse electromagnetic inhomogeneous medium problem. Find a tangential vector field $g \in L^2(\Omega)$, $m \in C_0^{1,\alpha}(B)$, and $(F, G) \in C^1(\bar{B})^2$ a solution of (1.7) in B that minimize the functional

$$M_4(g, F, G, m) := \|\mathbf{F}_e g - \tilde{E}_\infty\|_{L^2(\Omega)} + \|\hat{x} \times (F - \tilde{E} - E_g^i)\|_{L^2(\partial B)}$$
$$+ \|\hat{x} \times \operatorname{curl}(F - \tilde{E} - E_g^i)\|_{L^2(\partial B)}.$$

In the case of the acoustic inhomogeneous medium problem w and m are related by the Lippmann-Schwinger equation with incident field v and in the electromagnetic inhomogeneous medium problem F and m are related by the vector Lippmann-Schwinger equation (cf. [3]) with incident electric field E_g^i.

If κ is not an eigenvalue (as discussed in the opening paragraph), the infimum of each of these cost functionals is zero for exact far field data, and a solution of the inverse scattering problem is the limit of an appropriate minimizing sequence (cf. [3]). The same holds for the inhomogeneous medium problem for all $\kappa > 0$ in the case that n is not always real valued, i.e., $\operatorname{Im}(n(x)) \neq 0$ for some $x \in \mathbb{R}^3$ (cf. [11], whose methods also apply to the scalar case).

From the point of view of the numerical construction of solutions to the inverse scattering problem, it is highly desirable that for exact far field data the above cost functionals have an infimum equal to zero for all positive values of the wave number κ. Unfortunately, if κ is an eigenvalue this is not the case for each of the above cost functionals. This is due to the fact that for given far field patterns u_∞ and E_∞, there exist surfaces Γ and refractive indices n such that the integral operators $\mathbf{F}_a$ and $\mathbf{F}_e$ are not injective for a discrete set of κ values, and for such values of κ the functions $\tilde{u}_\infty$ and $\tilde{E}_\infty$ are in general not in the closure of the range of the operators $\mathbf{F}_a$ and $\mathbf{F}_e$, respectively (and hence the infimum of the cost functionals is not zero). The aim of this paper is to remedy this defect by replacing the kernel in the definition of $\mathbf{F}_a$ and $\mathbf{F}_e$ by $u_\infty - u_\infty^0$ and $E_\infty - E_\infty^0$, respectively, where u_∞^0 and E_∞^0 are the far field patterns of a surface potential. This change will of course necessitate changing the remaining terms in the cost functional in order to have the infimum be equal to zero.

To be more precise, consider the inverse acoustic obstacle problem, and let u^0 be the surface potential

$$u^0(x, d) := \frac{1}{4\pi} \int_{|y|=a} \varphi(y, d) \frac{\exp(i\kappa|x-y|)}{|x-y|} \, ds(y),$$

where a is such that if D is the unknown scattering obstacle, then it is known a priori that $\{x \in \mathbb{R}^3 : |x| \leqq a\} \subset D$. Then u^0 has the far field pattern

$$u_\infty^0(\hat{x}, d) := \frac{a^2}{4\pi} \int_\Omega \varphi(ay, d) \exp(-i\kappa a \hat{x} \cdot y) \, ds(y).$$

Now consider the operator $\mathbf{F}$ defined by

$$(1.8)\qquad \begin{aligned}(\mathbf{F}g)(\hat{x}) &:= \int_\Omega [u_\infty(\hat{x}, d) - u_\infty^0(\hat{x}, d)]g(d)\,ds(d)\\ &= \int_\Omega u_\infty(\hat{x}, d)g(d)\,ds(d) - a^2\int_\Omega (\mathbf{R}g)(d)\exp(-i\kappa a\hat{x}\cdot d)\,ds(d),\end{aligned}$$

where

$$(\mathbf{R}g)(y) := \frac{1}{4\pi}\int_\Omega \varphi(ay, d)g(d)\,ds(d), \qquad y \in \Omega.$$

Then by choosing a and φ appropriately or, more generally, choosing a and the operator $\mathbf{R}$ in (1.8) in an appropriate fashion, we will show that if in the definition of M_1 we replace the operator $\mathbf{F}_a$ by $\mathbf{F}$ and modify the second term, the resulting cost functional will have infimum equal to zero. More specifically, the integral equation $\mathbf{F}g = 0$ will have only the trivial solution $g = 0$, and $\tilde{u}_\infty$ will be in the closure of the range of $\mathbf{F}$. A similar analysis will also be done for the inverse acoustic inhomogeneous medium problem and the corresponding electromagnetic problems. These results remove the disadvantages of the dual space method referred to in the opening paragraph of this paper.

2. The inverse obstacle problem for acoustic waves. We denote by j_l the spherical Bessel function, by $h_l^{(1)}$, $h_l^{(2)}$ the spherical Hankel functions of the first and second kind of order l and by

$$\Phi(x, y) := \frac{1}{4\pi}\frac{\exp(i\kappa|x-y|)}{|x-y|}, \qquad x \neq y,$$

the fundamental solution to the Helmholtz equation. Let $a > 0$ be such that $\{x \in \mathbb{R}^3 \colon |x| \leqq a\} \subset D$ and $j_l(\kappa a) \neq 0$ for all $l \in \mathbb{N}_0$, i.e., $-\kappa^2$ is not an eigenvalue for the interior Dirichlet problem for the Laplace equation in $\{x \in \mathbb{R}^3 \colon |x| < a\}$. The condition that $\{x \in \mathbb{R}^3 \colon |x| \leqq a\} \subset D$ is only needed in the final step when we reformulate the inverse scattering problem as a constrained optimization problem. We denote by Y_l^k, $l = 0, 1, 2, \ldots, k = -l, \ldots, l$, an orthonormal basis of spherical harmonics on Ω. For $g \in L^2(\Omega)$,

$$g = \sum_{l=0}^{\infty}\sum_{k=-l}^{l} g_l^k Y_l^k,$$

we define

$$(2.1)\qquad \mathbf{R}g := \sum_{l=0}^{\infty}\sum_{k=-l}^{l} i^l g_l^k Y_l^k$$

and note that $\mathbf{R}\colon L^2(\Omega) \to L^2(\Omega)$ is a unitary operator.

We begin by defining the operator $\mathbf{F}$ for $g \in L^2(\Omega)$ by

$$(2.2)\qquad \begin{aligned}(\mathbf{F}g)(\hat{x}) := &\int_\Omega u_\infty(\hat{x}, d)g(d)\,ds(d)\\ &- a^2\int_\Omega (\mathbf{R}g)(d)\exp(-i\kappa a\hat{x}\cdot d)\,ds(d), \qquad \hat{x} \in \Omega.\end{aligned}$$

Our first aim is to prove the injectivity of the operator $\mathbf{F}$.

THEOREM 1. *Assume $\kappa > 0$. If $g \in L^2(\Omega)$ is a solution of* $\mathbf{F}g = 0$, *then* $g = 0$.

Proof. Let $g \in L^2(\Omega)$ be a solution of $\mathbf{F}g = 0$. We define

$$
\begin{aligned}
U^s(x) &:= \int_\Omega u^s(x, d) g(d)\, ds(d), \qquad x \in \bar{D}_e, \\
v(x) &:= \int_\Omega \exp(i\kappa x \cdot d) g(d)\, ds(d), \qquad x \in \mathbb{R}^3, \qquad (2.3)\\
u^0(x) &:= 4\pi \int_{|y|=a} \Phi(x, y)(\mathbf{R}g)\left(\frac{1}{a} y\right) ds(y), \qquad |x| > a.
\end{aligned}
$$

Then $U^s \in C^2(D_e) \cap C(\bar{D}_e)$ is a radiating solution of the Helmholtz equation in D_e, v is an entire solution to the Helmholtz equation, and $U^s(x) + v(x) = 0$ on ∂D. u^0 is also a radiating solution of the Helmholtz equation having the far field pattern

$$
\begin{aligned}
u^0_\infty(\hat{x}) &= a^2 \int_\Omega (\mathbf{R}g)(d) \exp(-i\kappa a \hat{x} \cdot d)\, ds(d) \\
&= \int_\Omega u_\infty(\hat{x}, d) g(d)\, ds(d) = U^s_\infty(\hat{x}), \qquad \hat{x} \in \Omega.
\end{aligned}
$$

Hence, if R is such that $\bar{D} \subset \{x \in \mathbb{R}^3 \colon |x| < R\}$ we have, by Rellich's lemma [2], [3], that $U^s(x) = u^0(x)$ for $|x| \geqq R$.

If $g \in L^2(\Omega)$ has the expansion

$$g = \sum_{l=0}^\infty \sum_{k=-l}^l g_l^k Y_l^k,$$

then

$$v(x) = 4\pi \sum_{l=0}^\infty \sum_{k=-l}^l g_l^k i^l j_l(\kappa|x|) Y_l^k(\hat{x}),$$

where the series converges uniformly with its derivatives on each compact set of $\mathbb{R}^3$ and

$$u^0(x) = 4\pi a^2 i\kappa \sum_{l=0}^\infty \sum_{k=-l}^l i^l g_l^k j_l(\kappa a) h_l^{(1)}(\kappa|x|) Y_l^k(\hat{x}),$$

where the series and its derivatives converge uniformly on each compact subset of $\{x \in \mathbb{R}^3 \colon |x| > a\}$. Using Green's theorem and $U^s + v = 0$ on ∂D we have

$$
\begin{aligned}
0 &= \int_{\partial D} \left[(U^s + v) \frac{\partial}{\partial \nu} \overline{(U^s + v)} - \overline{(U^s + v)} \frac{\partial}{\partial \nu} (U^s + v) \right] ds \\
&= \int_{|y|=R} \left[(u^0 + v) \frac{\partial}{\partial \nu} \overline{(u^0 + v)} - \overline{(u^0 + v)} \frac{\partial}{\partial \nu} (u^0 + v) \right] ds \\
&= \int_{|y|=R} \left(u^0 \frac{\partial}{\partial \nu} \overline{u^0} - \overline{u^0} \frac{\partial}{\partial \nu} u^0 \right) ds + 2i \operatorname{Im} \left\{ \int_{|y|=R} \left(u^0 \frac{\partial}{\partial \nu} \bar{v} - \bar{v} \frac{\partial}{\partial \nu} u^0 \right) ds \right\}.
\end{aligned}
$$

But

$$
\begin{aligned}
&\int_{|y|=R} \left(u^0 \frac{\partial}{\partial \nu} \overline{u^0} - \overline{u^0} \frac{\partial}{\partial \nu} u^0 \right) ds \\
&\quad = (4\pi a^2 \kappa)^2 \sum_{l=0}^\infty \sum_{k=-l}^l |g_l^k|^2 j_l(\kappa a)^2 (h_l^{(1)}(\kappa R) h_l^{(2)\prime}(\kappa R) - h_l^{(2)}(\kappa R) h_l^{(1)\prime}(\kappa R)) \kappa R^2 \\
&\quad = (4\pi a^2)^2 (-2i\kappa) \sum_{l=0}^\infty \sum_{k=-l}^l |g_l^k|^2 j_l(\kappa a)^2
\end{aligned}
$$

and

$$\int_{|y|=R}\left(u^0\frac{\partial}{\partial\nu}\bar{v}-\bar{v}\frac{\partial}{\partial\nu}u^0\right)ds$$

$$=(4\pi)^2a^2i\kappa\sum_{l=0}^{\infty}\sum_{k=-l}^{l}|g_l^k|^2j_l(\kappa a)(h_l^{(1)}(\kappa R)j_l'(\kappa R)-j_l(\kappa R)h_l^{(1)\prime}(\kappa R))\kappa R^2$$

$$=(4\pi)^2a^2\sum_{l=0}^{\infty}\sum_{k=-l}^{l}|g_l^k|^2j_l(\kappa a),$$

where we have used the Wronskian identities

$$W(h_l^{(1)},h_l^{(2)})(t)=h_l^{(1)}(t)h_l^{(2)\prime}(t)-h_l^{(2)}(t)h_l^{(1)\prime}(t)=\frac{-2i}{t^2},$$

$$W(h_l^{(1)},j_l)(t)=h_l^{(1)}(t)j_l'(t)-j_l(t)h_l^{(1)\prime}(t)=\frac{-i}{t^2}.$$

Hence we conclude that $g_l^k=0$ for all l, k, and thus $g=0$.

Remark 1. It is possible to get rid of the assumption $j_l(\kappa a)\neq 0$ for all $l\in\mathbb{N}_0$ by using

$$u^0(x)=4\pi\int_{|y|=a}\left\{\frac{\partial\Phi}{\partial\nu(y)}(x,y)-i\Phi(x,y)\right\}(\mathbf{R}g)\left(\frac{1}{a}y\right)ds(y),\qquad |x|>a,$$

and perturbing the kernel $u_\infty(\hat{x},d)$ by the corresponding far field pattern. In this case there are no longer any restrictions on a except that $a>0$ and $\{x\in\mathbb{R}^3\colon |x|\leqq a\}\subset D$.

Remark 2. In two dimensions with $d=(d_1,d_2)$ we have

$$g(d)=\sum_{l=-\infty}^{\infty}g_l(d_1+id_2)^l$$

and

$$(\mathbf{R}g)(d)=\sum_{l=-\infty}^{\infty}i^lg_l(d_1+id_2)^l=\sum_{l=-\infty}^{\infty}g_l(-d_2+id_1)^l.$$

Then

$$\int_\Omega\exp(-i\kappa a\hat{x}\cdot d)(\mathbf{R}g)(d)\,ds(d)=\int_\Omega\exp(i\kappa a\hat{x}\cdot d^*)g(d)\,ds(d),$$

where $d^*=(-d_2,d_1)$ has a simple geometrical interpretation.

The next lemma estimates the decay of the spherical Bessel function as $l\to\infty$.

LEMMA 1. For a fixed $t>0$ there exist positive constants $M>0$ and $C>0$ such that

$$j_l(t)\geqq C\frac{t^l}{1\cdot3\cdots(2l+1)}$$

for all $l\geqq M$.

Proof. Choose $M>t^2$. For $l\geqq M$ the sequence

$$a_p:=\left(\frac{t^2}{2}\right)^p\frac{1}{p!1\cdot3\cdots(2l+2p+1)},\qquad p\in\mathbb{N}_0,$$

is monotonically decreasing. Then we have for $l \geqq M$ that

$$j_l(t) = \sum_{p=0}^{\infty} \frac{(-1)^p t^{l+2p}}{2^p p! 1 \cdot 3 \cdots (2l+2p+1)}$$

$$\geqq \frac{t^l}{1 \cdot 3 \cdots (2l+1)} \left(1 - \frac{t^2}{2(2l+3)}\right) \geqq C \frac{t^l}{1 \cdot 3 \cdots (2l+1)}$$

for some constant $C > 0$.

The next theorem is a denseness theorem that is decisive in proving that the functional we shall replace M_1 with has infimum equal to zero for all $\kappa > 0$ (see the Introduction for the definition of M_1).

THEOREM 2. *Assume $\kappa > 0$. Then the linear space*

$$\operatorname{span}\{(j_l(\kappa|x|) + ia^2\kappa j_l(\kappa a) h_l^{(1)}(\kappa|x|))\, Y_l^k(\hat{x}) \colon l \in \mathbb{N}_0,\ k = -l, \ldots, l\}$$

is dense in $L^2(\partial D)$.

Proof. For $x, y \in \mathbb{R}^3$ we define χ by

$$\chi(x, y) = \frac{1}{a^2} \sum_{l=0}^{\infty} \sum_{k=-l}^{l} \frac{1}{j_l(\kappa a)} j_l(\kappa|x|) j_l(\kappa|y|)\, Y_l^k(\hat{x}) \overline{Y_l^k(\hat{y})}. \tag{2.4}$$

By Lemma 1 the series converges together with its derivatives uniformly on each compact subset of $\mathbb{R}^3 \times \mathbb{R}^3$. Assume $g \in L^2(\partial D)$ and

$$\int_{\partial D} \overline{g(x)}\, (j_l(\kappa|x|) + ia^2\kappa j_l(\kappa a) h_l^{(1)}(\kappa|x|))\, Y_l^k(\hat{x})\, ds(x) = 0$$

for all $l \in \mathbb{N}_0$, $k = -l, \ldots, l$. We must show $g = 0$. Define

$$u(x) := \int_{\partial D} \overline{g(y)} \{\Phi(x, y) + \chi(x, y)\}\, ds(y), \qquad x \in \mathbb{R}^3 \backslash \partial D.$$

Then $\Delta u + \kappa^2 u = 0$ in $\mathbb{R}^3 \backslash \partial D$ and if $r < a$ we have

$$\begin{aligned}
\int_{\Omega} u(r\hat{x})\, Y_l^k(\hat{x})\, ds(\hat{x}) &= \int_{\partial D} \overline{g(y)} \int_{\Omega} \{\Phi(r\hat{x}, y) + \chi(r\hat{x}, y)\} Y_l^k(\hat{x})\, ds(\hat{x})\, ds(y) \\
&= \int_{\partial D} \overline{g(y)} \Big\{ i\kappa h_l^{(1)}(\kappa|y|) j_l(\kappa r) \\
&\qquad + \frac{1}{a^2 j_l(\kappa a)} j_l(\kappa|y|) j_l(\kappa r) \Big\} Y_l^k(\hat{y})\, ds(y) \\
&= \frac{j_l(\kappa r)}{a^2 j_l(\kappa a)} \int_{\partial D} \overline{g(y)} \{j_l(\kappa|y|) + ia^2\kappa j_l(\kappa a) h_l^{(1)}(\kappa|y|)\} Y_l^k(\hat{y})\, ds(y) \\
&= 0
\end{aligned}$$

for all $l \in \mathbb{N}_0$, $k = -l, \ldots, l$. Hence $u = 0$ in a small ball with center at the origin and by unique continuation [2], [3], $u = 0$ in D.

For R such that $\bar{D} \subset \{x \in \mathbb{R}^3 \colon |x| < R\}$ we have $u(R\hat{x}) = u_1(R\hat{x}) + u_2(R\hat{x})$, where

$$\begin{aligned}
u_1(R\hat{x}) &:= \int_{\partial D} \Phi(R\hat{x}, y) \overline{g(y)}\, ds(y) \\
&= \sum_{l=0}^{\infty} \sum_{k=-l}^{l} \alpha_l^k h_l^{(1)}(\kappa R)\, Y_l^k(\hat{x})
\end{aligned}$$

and

$$u_2(R\hat{x}) := \int_{\partial D} \chi(R\hat{x}, y)\overline{g(y)}\,ds(y)$$

$$= \sum_{l=0}^{\infty} \sum_{k=-l}^{l} \frac{1}{a^2 i\kappa j_l(\kappa a)}\, \alpha_l^k j_l(\kappa R)\, Y_l^k(\hat{x})$$

with

$$\alpha_l^k = \int_{\partial D} i\kappa j_l(\kappa|y|)\overline{Y_l^k(\hat{y})g(y)}\,ds(y).$$

The series expansions for u_1 and u_2 converge uniformly together with their derivatives on $|x| = R$.

Using the fact that $u = 0$ in D and the continuity properties of single layer potentials with L^2 densities (see [12]) we conclude from Green's theorem that

$$0 = \int_{\partial D} \left\{ u \frac{\partial}{\partial \nu} \bar{u} - \bar{u} \frac{\partial}{\partial \nu} u \right\} ds$$

$$= \int_{|y|=R} \left\{ (u_1 + u_2) \frac{\partial}{\partial \nu} \overline{(u_1 + u_2)} - \overline{(u_1 + u_2)} \frac{\partial}{\partial \nu} (u_1 + u_2) \right\} ds$$

$$= \int_{|y|=R} \left\{ u_1 \frac{\partial}{\partial \nu} \overline{u_1} - \overline{u_1} \frac{\partial}{\partial \nu} u_1 \right\} ds + 2i \operatorname{Im} \left\{ \int_{|y|=R} \left\{ u_1 \frac{\partial}{\partial \nu} \overline{u_2} - \overline{u_2} \frac{\partial}{\partial \nu} u_1 \right\} ds \right\}.$$

But

$$\int_{|y|=R} \left\{ u_1 \frac{\partial}{\partial \nu} \overline{u_1} - \overline{u_1} \frac{\partial}{\partial \nu} u_1 \right\} ds$$

$$= \sum_{l=0}^{\infty} \sum_{k=-l}^{l} |\alpha_l^k|^2 \kappa R^2 (h_l^{(1)}(\kappa R) h_l^{(2)\prime}(\kappa R) - h_l^{(2)}(\kappa R) h_l^{(1)\prime}(\kappa R))$$

$$= \frac{-2i}{\kappa} \sum_{l=0}^{\infty} \sum_{k=-l}^{l} |\alpha_l^k|^2$$

and

$$\int_{|y|=R} \left\{ u_1 \frac{\partial}{\partial \nu} \overline{u_2} - \overline{u_2} \frac{\partial}{\partial \nu} u_1 \right\} ds$$

$$= -\frac{1}{a^2 i\kappa} \sum_{l=0}^{\infty} \sum_{k=-l}^{l} |\alpha_l^k|^2 \kappa R^2 \frac{1}{j_l(\kappa a)} (h_l^{(1)}(\kappa R) j_l'(\kappa R) - j_l(\kappa R) h_l^{(1)\prime}(\kappa R))$$

$$= \frac{1}{a^2 \kappa^2} \sum_{l=0}^{\infty} \sum_{k=-l}^{l} |\alpha_l^k|^2 \frac{1}{j_l(\kappa a)}.$$

Hence $\alpha_l^k = 0$ for all $l \in \mathbb{N}_0$, $k = -l, \ldots, l$, and thus $u = 0$ in D_e. We can now conclude from the jump relation for the normal derivative of the single layer potential that $g = 0$. The proof is finished.

If we now let $\tilde{u} \in C^2(\mathbb{R}^3 \setminus \{0\})$ be a radiating solution of the Helmholtz equation with the far field pattern $\tilde{u}_\infty$, we can choose for a given $\varepsilon > 0$ constants a_l^k, $l = 0, \dots, N$, $k = -l, \dots, l$ such that

$$\left\| \tilde{u} + 4\pi \sum_{l=0}^{N} \sum_{k=-l}^{l} a_l^k i^l [j_l(\kappa|\cdot|) + ia^2\kappa j_l(\kappa a) h_l^{(1)}(\kappa|\cdot|)] Y_l^k \right\|_{L^2(\partial D)} \leqq \varepsilon.$$

Define

$$\begin{aligned} g_N &:= \sum_{l=0}^{N} \sum_{k=-l}^{l} a_l^k Y_l^k, \\ v_N(x) &:= \int_\Omega \exp(i\kappa x \cdot d) g_N(d)\, ds(d) \\ &= 4\pi \sum_{l=0}^{N} \sum_{k=-l}^{l} a_l^k i^l j_l(\kappa|x|) Y_l^k(\hat{x}), \qquad x \in \mathbb{R}^3, \\ U_N^s(x) &:= \int_\Omega u^s(x, d) g_N(d)\, ds(d), \qquad x \in \bar{D}_e, \end{aligned} \tag{2.5}$$

and

$$u_N^0(x) := i4\pi\kappa a^2 \sum_{l=0}^{N} \sum_{k=-l}^{l} a_l^k i^l j_l(\kappa a) h_l^{(1)}(\kappa|x|) Y_l^k(\hat{x}), \qquad |x| > 0.$$

Then using $U_N^s + v_N = 0$ on ∂D we have that

$$\|\tilde{u} - (U_N^s - u_N^0)\|_{L^2(\partial D)} \leqq \varepsilon,$$

and since

$$u_N^0(x) = 4\pi a^2 \int_\Omega \Phi(x, ad)(\mathbf{R} g_N)(d)\, ds(d), \qquad |x| > a,$$

we have

$$\sup_{\hat{x} \in \Omega} |\tilde{u}_\infty(\hat{x}) - (U_{N,\infty}^s(\hat{x}) - u_{N,\infty}^0(\hat{x}))| \leqq M\varepsilon$$

for some positive constant M, i.e.,

$$|(\mathbf{F} g_N)(\hat{x}) - \tilde{u}_\infty(\hat{x})| \leqq M\varepsilon$$

for all $\hat{x} \in \Omega$. Hence, if v and u^0 are defined by (2.3), u_∞ is the exact far field data, and S is the set of C^2 surfaces bounding a domain D containing the origin such that D_e is connected, the infimum of the functional

$$M_1(g, \Gamma) := \|\mathbf{F} g - \tilde{u}_\infty\|_{L^2(\Omega)} + \|\tilde{u} + v + u^0\|_{L^2(\Gamma)}$$

for $g \in L^2(\Omega)$, $\Gamma \in S$, is zero for all $\kappa > 0$, and the solution of the inverse scattering problem is the limit of an appropriate minimizing sequence. The aim of this section of our paper is now accomplished.

3. The inverse inhomogeneous medium problem for acoustic waves. We define the operator $\mathbf{F}$ as in (2.2), where $u_\infty(\hat{x}, d)$ is now to be understood as the far field pattern of the inhomogeneous medium problem corresponding to an incident plane wave propagating in direction d. In this section, $a > 0$ is only required to satisfy $j_l(\kappa a) \neq 0$ for all $l \in \mathbb{N}_0$. We again first prove that the operator $\mathbf{F}$ is injective.

THEOREM 3. *Assume $\kappa > 0$. If $g \in L^2(\Omega)$ is a solution of $\mathbf{F}g = 0$, then $g = 0$.*

Proof. The proof is almost identical to the proof of Theorem 1. The equality

$$0 = \int_{|y|=R} \left[(U^s + v) \frac{\partial}{\partial \nu} \overline{(U^s + v)} - \overline{(U^s + v)} \frac{\partial}{\partial \nu} (U^s + v) \right] ds$$

follows by Green's theorem and the fact that $w = (U^s + v)$ is a solution of $\Delta w + \kappa^2 n w = 0$ in $\mathbb{R}^3$, where n is real valued.

In the next theorem we assume that $\operatorname{supp}(1 - n) \subset B := \{x \in \mathbb{R}^3 \colon |x| < R\}$, where $R > a$, and define

$$X := \operatorname{span} \{(j_l(\kappa|x|) + ia^2 \kappa j_l(\kappa a) h_l^{(1)}(\kappa|x|)) Y_l^k(\hat{x}) \colon l \in \mathbb{N}_0, k = -l, \ldots, l\},$$

$$Y := \{w \in C^2(\bar{B}) \colon \Delta w + \kappa^2 n w = 0 \text{ in } B\},$$

$$W := \left\{ \left(h - w, \frac{\partial}{\partial r}(h - w) \right) \colon h \in X, w \in Y \right\} \subset L^2(\partial B) \times L^2(\partial B).$$

THEOREM 4. *Assume $\kappa > 0$. Then W is dense in $L^2(\partial B) \times L^2(\partial B)$.*

Proof. Suppose a, $b \in L^2(\partial B)$ satisfy

$$\int_{\partial B} \left[(h - w)\bar{a} + \frac{\partial}{\partial r}(h - w)\bar{b} \right] ds = 0 \tag{3.1}$$

for all $h \in X$, $w \in Y$. We must show $a = b = 0$. To this end, for χ given by (2.4) we define v by

$$v(x) := \int_{\partial B} \left[\overline{a(y)} \{\Phi(x, y) + \chi(x, y)\} + \overline{b(y)} \frac{\partial}{\partial |y|} \{\Phi(x, y) + \chi(x, y)\} \right] ds(y)$$

for $x \in \mathbb{R}^3 \setminus \partial B$. We compute, for $r < R$,

$$\begin{aligned} &\int_{\Omega} v(r\hat{x}) Y_l^k(\hat{x}) \, ds(\hat{x}) \\ &= \int_{\partial B} \overline{a(y)} \left\{ i\kappa j_l(\kappa r) h_l^{(1)}(\kappa|y|) + \frac{1}{a^2 j_l(\kappa a)} j_l(\kappa r) j_l(\kappa|y|) \right\} Y_l^k(\hat{y}) \, ds(y) \\ &+ \int_{\partial B} \overline{b(y)} \frac{\partial}{\partial |y|} \left\{ i\kappa j_l(\kappa r) h_l^{(1)}(\kappa|y|) + \frac{1}{a^2 j_l(\kappa a)} j_l(\kappa r) j_l(\kappa|y|) \right\} Y_l^k(\hat{y}) \, ds(y) = 0 \end{aligned}$$

for all $l \in \mathbb{N}_0$, $k = -l, \ldots, l$. Therefore, $v = 0$ in B, and we are able to conclude by the L^2 jump relations for surface potentials [12] that $v_+ := v|_{\mathbb{R}^3 \setminus \bar{B}}$ satisfies

$$v_+ = \bar{b} \quad \text{and} \quad \frac{\partial}{\partial r} v_+ = -\bar{a} \tag{3.2}$$

on ∂B. From (3.1) we now have that, for all $w \in Y$,

$$\int_{\partial B} \left(\frac{\partial v_+}{\partial r} w - v_+ \frac{\partial w}{\partial r} \right) ds = 0. \tag{3.3}$$

We want to show through a suitable choice of w that $v_+ = 0$. The theorem then follows by (3.2).

We define v^s and v^0 by

$$v^s(x) := \int_{\partial B} \left[\overline{a(y)} \Phi(x, y) + \overline{b(y)} \frac{\partial \Phi}{\partial |y|}(x, y) \right] ds(y), \qquad |x| > R,$$

$$v^0(x) := \int_{\partial B} \left[\overline{a(y)} \chi(x, y) + \overline{b(y)} \frac{\partial \chi}{\partial |y|}(x, y) \right] ds(y), \qquad x \in \mathbb{R}^3.$$

Then $v(x) = v^0(x) + v^s(x)$ for $|x| > R$, and

$$v^0(r\hat{x}) = 4\pi \sum_{l=0}^{\infty} \sum_{k=-l}^{l} i^l \alpha_l^k j_l(\kappa r) Y_l^k(\hat{x}), \qquad r \geqq 0,$$

$$v^s(r\hat{x}) = 4\pi \sum_{l=0}^{\infty} \sum_{k=-l}^{l} i^l \alpha_l^k (i\kappa a^2) j_l(\kappa a) h_l^{(1)}(\kappa r) Y_l^k(\hat{x}), \qquad r > R,$$

where

$$\alpha_l^k = \frac{(-i)^l}{4\pi a^2 j_l(\kappa a)} \int_{\partial B} [j_l(\kappa |y|) \overline{Y_l^k(\hat{y}) a(y)} + \kappa j_l'(\kappa |y|) \overline{Y_l^k(\hat{y}) b(y)}]\, ds(y),$$

$l \in \mathbb{N}_0$, $k = -l, \ldots, l$.

Since $\sum_{l=0}^{N} \sum_{k=-l}^{l} |\alpha_l^k|^2$ is possibly divergent for $N \to \infty$ we define

$$g_N := \sum_{l=0}^{N} \sum_{k=-l}^{l} \alpha_l^k Y_l^k$$

$$v_N^0(x) := \int_{\Omega} \exp(i\kappa x \cdot d) g_N(d)\, ds(d)$$

$$= 4\pi \sum_{l=0}^{N} \sum_{k=-l}^{l} i^l \alpha_l^k j_l(\kappa |x|) Y_l^k(\hat{x})$$

and note that $v_N^0(x) \to v^0(x)$, $N \to \infty$, together with its derivatives uniformly on $\bar{B}$ since v^0 is an entire solution of the Helmholtz equation and v_N^0 are the partial Fourier sums of v^0. From the Lippmann-Schwinger equation (see [3, Chap. 8]) we see that the solutions U_N of the inhomogeneous medium problem corresponding to the incident wave v_N^0 converge together with their first and second derivatives uniformly on any bounded subset of $\mathbb{R}^3$ to the solution U corresponding to the incident wave v^0, $U_N \to U$, $N \to \infty$. Again, using the Lippmann-Schwinger equation, we can conclude that the scattered fields $U_N^s := U_N - v_N$ converge together with their derivatives to $U^s := U - v^0$ uniformly on $\mathbb{R}^3$, $U_N^s \to U^s$, $N \to \infty$.

We now want to show that $U^s(x) = v^s(x)$ for $|x| > R$. We know that

$$u_\infty(\hat{x}, d) = -\frac{\kappa^2}{4\pi} \int_B \exp(-i\kappa y \cdot \hat{x}) m(y) u(y, d)\, dy, \qquad \hat{x} \in \Omega, \tag{3.4}$$

and using (3.3) and Green's theorem we have that

$$\begin{aligned} &\int_{\partial B} \left[v^s(y) \frac{\partial}{\partial |y|} \exp(-i\kappa y \cdot \hat{x}) - \frac{\partial v^s}{\partial |y|}(y) \exp(-i\kappa y \cdot \hat{x}) \right] ds(y) \\ &= \int_{\partial B} \left[v^s(y) \frac{\partial}{\partial |y|} (\exp(-i\kappa y \cdot \hat{x}) + u^s(y, -\hat{x})) \right. \\ &\qquad \left. - \frac{\partial v^s}{\partial |y|}(y) (\exp(-i\kappa y \cdot \hat{x}) + u^s(y, -\hat{x})) \right] ds(y) \\ &= -\int_{\partial B} \left[v^0(y) \frac{\partial}{\partial |y|} u(y, -\hat{x}) - \frac{\partial v^0}{\partial |y|}(y) u(y, -\hat{x}) \right] ds(y) \\ &= -\kappa^2 \int_B m(y) v^0(y) u(y, -\hat{x})\, dy. \end{aligned} \tag{3.5}$$

From the reciprocity theorem [3, Thm. 8.8], (3.4), and (3.5) we now see that if U^s_∞ is the far field pattern of U^s, then

$$\begin{aligned}
U^s_\infty(\hat{x}) &= \lim_{N\to\infty} \int_\Omega g_N(d)u_\infty(\hat{x}, d)\, ds(d) \\
&= \lim_{N\to\infty} \int_\Omega g_N(d)u_\infty(-d, -\hat{x})\, ds(d) \\
&= -\frac{\kappa^2}{4\pi} \lim_{N\to\infty} \int_\Omega g_N(d) \int_B \exp(i\kappa d\cdot y)m(y)u(y, -\hat{x})\, dy\, ds(d) \\
&= -\frac{\kappa^2}{4\pi} \lim_{N\to\infty} \int_B v^0_N(y)m(y)u(y, -\hat{x})\, dy \\
&= -\frac{\kappa^2}{4\pi} \int_B v^0(y)m(y)u(y, -\hat{x})\, dy \\
&= \frac{1}{4\pi} \int_{\partial B} \left[v^s(y)\frac{\partial}{\partial |y|}\exp(-i\kappa y\cdot\hat{x}) - \frac{\partial v^s}{\partial |y|}(y)\exp(-i\kappa y\cdot \hat{x})\right] ds(y) \\
&= v^s_\infty(\hat{x}),
\end{aligned}$$

and hence by Rellich's lemma [2], [3], $U^s(x) = v^s(x)$ for $|x| > R$.

Finally, we compute for $R_1 > R$ that

$$\begin{aligned}
0 &= \int_{|y|=R_1} \left[(v^0+v^s)\frac{\partial}{\partial r}\overline{(v^0+v^s)} - \overline{(v^0+v^s)}\frac{\partial}{\partial r}(v^0+v^s)\right] ds \\
&= \int_{|y|=R_1} \left[v^s\frac{\partial}{\partial r}\overline{v^s} - \overline{v^s}\frac{\partial}{\partial r}v^s\right] ds + 2i \operatorname{Im}\left\{ \int_{|y|=R_1} \left[v^s\frac{\partial}{\partial r}\overline{v^0} - \overline{v^0}\frac{\partial}{\partial r}v^s\right] ds \right\} \\
&= (4\pi\kappa a^2)^2 \sum_{l=0}^{\infty}\sum_{k=-l}^{l} j_l(\kappa a)^2|\alpha_l^k|^2 W(h_l^{(1)}, h_l^{(2)})(\kappa R_1)\kappa R_1^2 \\
&= -2i\kappa(4\pi a^2)^2 \sum_{l=0}^{\infty}\sum_{k=-l}^{l} j_l(\kappa a)^2|\alpha_l^k|^2,
\end{aligned}$$

where W denotes the Wronskian, and hence $\alpha_l^k = 0$ for all $l \in \mathbb{N}_0$, $k = -l, \dots, l$. We can now conclude that $v_+ = 0$, and our proof is finished.

Now let $\tilde{u} \in C^2(\mathbb{R}^3\backslash\{0\})$ be a radiating solution of the Helmholtz equation with far field pattern $\tilde{u}_\infty$. If we set $m := 1 - n$ then, according to Theorem 4, for a given $\varepsilon > 0$ we can find the complex numbers a_l^k, $l = 0, \dots, N$, $k = -l, \dots, l$, and $w \in Y$ such that

$$\begin{aligned}
&\left\| w - \tilde{u} - 4\pi \sum_{l=0}^{N}\sum_{k=-l}^{l} a_l^k i^l [j_l(\kappa|\cdot|) + ia^2\kappa j_l(\kappa a)h_l^{(1)}(\kappa|\cdot|)]Y_l^k \right\|_{L^2(\partial B)} \\
&\quad + \left\| \frac{\partial}{\partial r}\left(w - \tilde{u} - 4\pi \sum_{l=0}^{N}\sum_{k=-l}^{l} a_l^k i^l [j_l(\kappa|\cdot|) + ia^2\kappa j_l(\kappa a)h_l^{(1)}(\kappa|\cdot|)]Y_l^k \right) \right\|_{L^2(\partial B)} \leqq \varepsilon.
\end{aligned} \tag{3.6}$$

We define

$$v^0(x) := \int_{\partial B} \left(\Phi(x, y) \frac{\partial w}{\partial |y|}(y) - \frac{\partial \Phi}{\partial |y|}(x, y) w(y) \right) ds(y), \qquad x \in B,$$

$$w^s(x) := -\int_B \Phi(x, y)(\Delta w(y) + \kappa^2 w(y))\, dy, \qquad x \in \mathbb{R}^3,$$

$$g_N := \sum_{l=0}^{N} \sum_{k=-l}^{l} a_l^k Y_l^k,$$

$$v_N^0(x) := \int_\Omega \exp(i\kappa x \cdot d) g_N(d)\, ds(d)$$

$$= 4\pi \sum_{l=0}^{N} \sum_{k=-l}^{l} a_l^k i^l j_l(\kappa |x|) Y_l^k(\hat{x}), \qquad x \in \mathbb{R}^3,$$

$$u_N^s(x) := \int_\Omega u^s(x, d) g_N(d)\, ds(d), \qquad x \in \mathbb{R}^3.$$

From Green's representation theorem we can conclude that $w = v^0 + w^s$. Again, applying Green's representation theorem to

$$w - \tilde{u} - 4\pi \sum_{l=0}^{N} \sum_{k=-l}^{l} a_l^k i^l [j_l(\kappa |\cdot|) + ia^2 \kappa j_l(\kappa a) h_l^{(1)}(\kappa |\cdot|)] Y_l^k \tag{3.7}$$

in B and using (3.6), we have for $R_1 < R$ such that $\operatorname{supp}(m) \subset \{x \in \mathbb{R}^3 : |x| < R_1\}$ that

$$\sup_{|x| \leqq R_1} |v^0(x) - v_N^0(x)| \leqq C_1 \varepsilon,$$

and hence from the Lippmann-Schwinger equation

$$\sup_{x \in \mathbb{R}^3} |w^s(x) - u_N^s(x)| \leqq C_2 \varepsilon$$

for positive constants C_1, C_2. By applying Green's representation thereorem to the function (3.7) in $\mathbb{R}^3 \backslash \bar{B}$ and using the L^2 jump relations for surface potentials [12] we get

$$\left\| w^s - \tilde{u} - 4\pi \sum_{l=0}^{N} \sum_{k=-l}^{l} a_l^k i^l ia^2 \kappa j_l(\kappa a) h_l^{(1)}(\kappa |\cdot|) Y_l^k \right\|_{L^2(\partial B)} \leqq C_3 \varepsilon$$

for a positive constant C_3. We can conclude that, for $R_2 > R$,

$$\sup_{|x| = R_2} \left| u_N^s(x) - \tilde{u}(x) - 4\pi \sum_{l=0}^{N} \sum_{k=-l}^{l} a_l^k i^l ia^2 \kappa j_l(\kappa a) h_l^{(1)}(\kappa |x|) Y_l^k(\hat{x}) \right| \leqq C_4 \varepsilon$$

for a positive constant C_4, and hence

$$\sup_{\hat{x} \in \Omega} \left| \int_\Omega u_\infty(\hat{x}, d) g_N(d)\, ds(d) - a^2 \int_\Omega \exp(-i\kappa a \hat{x} \cdot d)(\mathbf{R} g_N)(d)\, ds(d) - \tilde{u}_\infty(\hat{x}) \right| \leqq C_5 \varepsilon$$

for a positive constant C_5. Therefore, if v and u^0 are defined by (2.3) and u_∞ is the exact far field data, the infimum of the functional

$$M_2(g, w, m) := \|\mathbf{F} g - \tilde{u}_\infty\|_{L^2(\Omega)} + \|w - \tilde{u} - v - u^0\|_{L^2(\partial B)} + \left\| \frac{\partial}{\partial r}(w - \tilde{u} - v - u^0) \right\|_{L^2(\partial B)}$$

for $g \in L^2(\Omega)$, $m \in C_0^{0,\alpha}(B)$, $w \in Y$ is zero for all $\kappa > 0$, and the solution of the inverse scattering problem is the limit of an appropriate minimizing sequence. This completes the task of this section.

4. The inverse obstacle problem for electromagnetic waves. In this section we shall consider the inverse obstacle problem for electromagnetic waves in a manner analogous to that of § 2 for acoustic waves. We begin by defining some additional notation to that given in the Introduction. Let $a>0$ be such that $\{x\in\mathbb{R}^3\colon |x|\leqq a\}\subset D$, and

$$j_l(\kappa a)(j_l(\kappa a)+\kappa a j_l'(\kappa a))\neq 0$$

for all $l\in\mathbb{N}$, i.e., κ is not an eigenvalue for the interior Dirichlet problem for Maxwell's equations in $\{x\in\mathbb{R}^3\colon |x|<a\}$. We denote by Y_l^k, $l=0,1,2,\dots,k=-l,\dots,l$, an orthonormal basis of spherical harmonics on Ω and by

$$U_l^k(d):=\frac{1}{(l(l+1))^{1/2}}\operatorname{Grad} Y_l^k(d),\qquad V_l^k(d):=d\times U_l^k(d),\quad d\in\Omega,$$

$l\in\mathbb{N}$, $k=-l,\dots,l$, with Grad denoting the surface gradient, an orthonormal basis of $T^2(\Omega)$, the square integrable tangential fields on Ω (see [3]). For $g\in T^2(\Omega)$,

$$g=\sum_{l=1}^{\infty}\sum_{k=-l}^{l} a_l^k U_l^k+b_l^k V_l^k, \tag{4.1}$$

we define

$$\mathbf{R}g:=\sum_{l=1}^{\infty}\sum_{k=-l}^{l} i^l a_l^k V_l^k+ii^l b_l^k U_l^k$$

and note that $\mathbf{R}\colon T^2(\Omega)\to T^2(\Omega)$ is a unitary operator. Furthermore, we define

$$M_l^k(x):=\operatorname{curl}\{xj_l(\kappa|x|)Y_l^k(\hat{x})\},\qquad x\in\mathbb{R}^3,$$

and

$$N_l^k(x):=\operatorname{curl}\{xh_l^{(1)}(\kappa|x|)Y_l^k(\hat{x})\},\qquad x\in\mathbb{R}^3\backslash\{0\}.$$

We compute (use the asymptotic form of (6.71) and (6.26) in [3])

$$\int_\Omega \exp(i\kappa x\cdot d)U_l^k(d)\,ds(d)=\frac{4\pi i^l}{(l(l+1))^{1/2}i\kappa}\operatorname{curl} M_l^k(x)$$

and

$$\int_\Omega \exp(i\kappa x\cdot d)V_l^k(d)\,ds(d)=\frac{-4\pi i^l}{(l(l+1))^{1/2}}M_l^k(x),\qquad x\in\mathbb{R}^3,$$

and note the relations

$$\int_{|y|=R}\nu\times N_l^k\cdot\operatorname{curl}\overline{N_{\tilde{l}}^{\tilde{k}}}\,ds=l(l+1)R^2h_l^{(1)}(\kappa R)\frac{1}{R}(h_l^{(2)}(\kappa R)+\kappa Rh_l^{(2)\prime}(\kappa R))\delta_{l\tilde{l}}\delta_{k\tilde{k}},$$

$$\int_{|y|=R}\nu\times\operatorname{curl} N_l^k\cdot\operatorname{curlcurl}\overline{N_{\tilde{l}}^{\tilde{k}}}\,ds$$
$$=-\kappa^2R^2l(l+1)h_l^{(2)}(\kappa R)\frac{1}{R}(h_l^{(1)}(\kappa R)+\kappa Rh_l^{(1)\prime}(\kappa R))\delta_{l\tilde{l}}\delta_{k\tilde{k}},$$

$$\int_{|y|=R}\nu\times N_l^k\cdot\overline{N_{\tilde{l}}^{\tilde{k}}}\,ds=\int_{|y|=R}\nu\times\operatorname{curl} N_l^k\cdot\operatorname{curl}\overline{N_{\tilde{l}}^{\tilde{k}}}\,ds=0$$

for all l, $\tilde{l}\in\mathbb{N}$, $k=-l,\dots,l$, $\tilde{k}=-\tilde{l},\dots,\tilde{l}$. There are similar relations if N_l^k is replaced by M_l^k, in which case $h_l^{(1)}$ has to be replaced by j_l.

Analogous to the case of acoustic waves, we begin by defining the operator $\mathbf{F}$ for $g\in T^2(\Omega)$ by

$$\begin{aligned}(\mathbf{F}g)(\hat{x}):=&\int_\Omega E_\infty(\hat{x},d,g(d))\,ds(d)\\ &-\frac{i\kappa a^2}{4\pi}\hat{x}\times\int_\Omega(\mathbf{R}g)(d)\exp(-i\kappa a\hat{x}\cdot d)\,ds(d),\qquad \hat{x}\in\Omega,\end{aligned}\tag{4.2}$$

where $E_\infty(\hat{x}, d, p)$ is the electric far field pattern corresponding to the incident field (1.1) being scattered by the obstacle D.

We first show that the operator F is injective.

THEOREM 5. *Assume $\kappa > 0$. If $g \in T^2(\Omega)$ is a solution of* $\mathbf{F}g = 0$, *then* $g = 0$.

Proof. Let $g \in T^2(\Omega)$ be a solution of $\mathbf{F}g = 0$. We define

$$
\begin{aligned}
E^s_g(x) &:= \int_\Omega E^s(x, d, g(d))\, ds(d), \qquad x \in \bar{D}_e,\\
E^i_g(x) &:= \int_\Omega E^i(x, d, g(d))\, ds(d)\\
&= i\kappa \int_\Omega \exp(i\kappa x\cdot d) g(d)\, ds(d), \qquad x \in \mathbb{R}^3,\\
E^0_g(x) &:= a^2 \operatorname{curl} \int_\Omega \Phi(x, ad)(\mathbf{R}g)(d)\, ds(d), \qquad |x| > a.
\end{aligned}
\tag{4.3}
$$

Since $\mathbf{F}g = 0$ we have that $E^0_{g,\infty} = E_{g,\infty}$ and hence by Rellich's lemma [2], [3], $E^0_g(x) = E^s_g(x)$ for $|x| > R$, where R is such that $\bar{D} \subset \{x \in \mathbb{R}^3 \colon |x| < R\}$. Using Green's theorem we have

$$
\begin{aligned}
0 &= \int_{\partial D} [\nu \times (E^i_g + E^s_g)\cdot \operatorname{curl} \overline{(E^i_g + E^s_g)} - \nu \times \overline{(E^i_g + E^s_g)}\cdot \operatorname{curl}(E^i_g + E^s_g)]\, ds\\
&= \int_{|y|=R} [\nu \times (E^i_g + E^0_g)\cdot \operatorname{curl} \overline{(E^i_g + E^0_g)} - \nu \times \overline{(E^i_g + E^0_g)}\cdot \operatorname{curl}(E^i_g + E^0_g)]\, ds\\
&= \int_{|y|=R} [\nu \times E^0_g \cdot \operatorname{curl} \overline{E^0_g} - \nu \times \overline{E^0_g}\cdot \operatorname{curl} E^0_g]\, ds\\
&\quad + 2i \operatorname{Im}\left\{\int_{|y|=R} [\nu \times E^0_g \cdot \operatorname{curl} \overline{E^i_g} - \nu \times \overline{E^i_g}\cdot \operatorname{curl} E^0_g]\, ds\right\}.
\end{aligned}
\tag{4.4}
$$

We will now calculate the series expansions of E^0_g and E^s_g and put them into (4.4). Let g have the expansion (4.1). Then

$$
\begin{aligned}
E^i_g(x) &= i\kappa \sum_{l=1}^{\infty} \sum_{k=-l}^{l} a^k_l \int_\Omega \exp(i\kappa x\cdot d) U^k_l(d)\, ds(d) + b^k_l \int_\Omega \exp(i\kappa x\cdot d) V^k_l(d)\, ds(d)\\
&= 4\pi \sum_{l=1}^{\infty} \sum_{k=-l}^{l} i^l a^k_l \frac{1}{[l(l+1)]^{1/2}} \operatorname{curl} M^k_l(x)\\
&\quad - 4\pi i\kappa \sum_{l=1}^{\infty} \sum_{k=-l}^{l} i^l b^k_l \frac{1}{[l(l+1)]^{1/2}} M^k_l(x), \qquad x \in \mathbb{R}^3,
\end{aligned}
$$

$$
\begin{aligned}
E^0_g(x) &= a^2 \sum_{l=1}^{\infty} \sum_{k=-l}^{l} \left\{ i^l a^k_l \operatorname{curl} \int_\Omega \Phi(x, ad) V^k_l(d)\, ds(d)\right.\\
&\qquad \left. + i^l b^k_l i \operatorname{curl} \int_\Omega \Phi(x, ad) U^k_l(d)\, ds(d)\right\}\\
&= a^2 \sum_{l=1}^{\infty} \sum_{k=-l}^{l} i^l a^k_l \left(-i\kappa \frac{1}{[l(l+1)]^{1/2}}\right) j_l(\kappa a) \operatorname{curl} N^k_l(x)\\
&\quad + a^2 \sum_{l=1}^{\infty} \sum_{k=-l}^{l} i^{(l+1)} b^k_l \left(i\kappa \frac{1}{[l(l+1)]^{1/2}}\right) \frac{1}{a} \{j_l(\kappa a) + \kappa a j_l'(\kappa a)\} N^k_l(x).
\end{aligned}
$$

Then we calculate for the first term in (4.4)

$$\int_{|y|=R} [\nu \times E_g^0 \cdot \operatorname{curl} \overline{E_g^0} - \nu \times \overline{E_g^0} \cdot \operatorname{curl} E_g^0]\, ds$$

$$= -2ia^4\kappa^3 \sum_{l=1}^{\infty} \sum_{k=-l}^{l} |a_l^k|^2 j_l(\kappa a)^2$$

$$-2ia^4\kappa \sum_{l=1}^{\infty} \sum_{k=-l}^{l} |b_l^k|^2 \left[\frac{1}{a}(j_l(\kappa a) + \kappa a j_l'(\kappa a))\right]^2,$$

and for the second term in (4.4),

$$\int_{|y|=R} [\nu \times E_g^0 \cdot \operatorname{curl} \overline{E_g^i} - \nu \times \overline{E_g^i} \cdot \operatorname{curl} E_g^0]\, ds$$

$$= -4\pi\kappa^2 a^2 \sum_{l=1}^{\infty} \sum_{k=-l}^{l} |a_l^k|^2 j_l(\kappa a)$$

$$-4\pi a^2 \kappa \sum_{l=1}^{\infty} \sum_{k=-l}^{l} |b_l^k|^2 \frac{1}{a}(j_l(\kappa a) + \kappa a j_l'(\kappa a)) \in \mathbb{R}.$$

Hence we conclude from (4.4) that $a_l^k = b_l^k = 0$ for all $l \in \mathbb{N}$, $k = -l, \dots, l$, i.e., $g = 0$. The theorem is now proved.

The next theorem is the analogous result to Theorem 2.

THEOREM 6. *Assume $\kappa > 0$. Then the linear space*

$$\operatorname{span}\{\nu \times (4\pi \operatorname{curl} M_l^k - ia^2\kappa j_l(\kappa a) \operatorname{curl} N_l^k),$$

$$\nu \times (4\pi M_l^k - ia\{j_l(\kappa a) + \kappa a j_l'(\kappa a)\} N_l^k),\, l \in \mathbb{N},\, k = -l, \dots, l\}$$

is dense in $T^2(\partial D)$.

Proof. We define for $x, y \in \mathbb{R}^3$ a matrix valued function $\chi(x, y)$ such that for $p \in \mathbb{C}^3$,

$$\chi(x, y)p = -4\pi\kappa \sum_{l=1}^{\infty} \frac{1}{l(l+1)} \frac{1}{a(j_l(\kappa a) + \kappa a j_l'(\kappa a))} \sum_{k=-l}^{l} M_l^k(y) \cdot p \operatorname{curl} \overline{M_l^k(x)}$$

(4.5)

$$-4\pi \sum_{l=1}^{\infty} \frac{1}{l(l+1)} \frac{1}{a^2 j_l(\kappa a)} \sum_{k=-l}^{l} \operatorname{curl} M_l^k(y) \cdot p \overline{M_l^k(x)}.$$

Similar arguments as in Theorem 2 show after some computation that the series converges uniformly along with its derivatives on compact subsets of $\mathbb{R}^3 \times \mathbb{R}^3$. Now let $g \in T^2(\partial D)$ be orthogonal to the linear space stated in the theorem. Define for $x \in \mathbb{R}^3 \backslash \partial D$,

$$E_g(x) := \operatorname{curl} \int_{\partial D} \Phi(x, y)\nu(y) \times \overline{g(y)}\, ds(y) + \int_{\partial D} \chi(x, y)(\nu(y) \times \overline{g(y)})\, ds(y).$$

For $r < a$ we have, using the vector addition theorem for the fundamental solution Φ [3, Thm. 6.27], that

$$\int_{\Omega} E_g(r\hat{x}) \cdot U_l^k(\hat{x})\, ds(\hat{x}) = \int_{\Omega} E_g(r\hat{x}) \cdot V_l^k(\hat{x})\, ds(\hat{x}) = 0$$

for all $l \in \mathbb{N}$, $k = -l, \dots, l$, and hence by unique continuation $E_g(x) = 0$ for $x \in D$. For R such that $\bar{D} \subset \{x \in \mathbb{R}^3 \colon |x| < R\}$ we again have from the vector addition theorem for

Φ that $E_g(R\hat{x}) = E_1(R\hat{x}) + E_2(R\hat{x})$ with

$$E_1(x) := \operatorname{curl} \int_{\partial D} \Phi(x, y)\nu(y) \times \overline{g(y)}\, ds(y)$$

$$= \sum_{l=1}^{\infty} \sum_{k=-l}^{l} \alpha_l^k N_l^k(x) + \beta_l^k \operatorname{curl} N_l^k(x),$$

$$E_2(x) := \int_{\partial D} \chi(x, y)(\nu(y) \times \overline{g(y)})\, ds(y)$$

$$= \sum_{l=1}^{\infty} \sum_{k=-l}^{l} -\frac{4\pi}{i\kappa a^2 j_l(\kappa a)} \alpha_l^k M_l^k(x) - \frac{4\pi\beta_l^k}{ia(j_l(\kappa a) + \kappa a j_l'(\kappa a))} \operatorname{curl} M_l^k(x),$$

where

$$\alpha_l^k = \frac{i\kappa}{l(l+1)} \int_{\partial D} (\nu(y) \times \overline{g(y)}) \cdot \operatorname{curl} \overline{M_l^k(y)}\, ds(y),$$

$$\beta_l^k = \frac{i\kappa}{l(l+1)} \int_{\partial D} (\nu(y) \times \overline{g(y)}) \cdot \overline{M_l^k(y)}\, ds(y).$$

By the continuity properties of vector potentials with L^2 densities (cf. [10], [11]) we have

$$0 = \int_{\partial D} \{\nu \times (E_1 + E_2) \cdot \operatorname{curl} \overline{(E_1 + E_2)} - \nu \times \overline{(E_1 + E_2)} \cdot \operatorname{curl} (E_1 + E_2)\}\, ds$$

$$= \int_{|y|=R} \{\nu \times E_1 \cdot \operatorname{curl} \overline{E_1} - \nu \times \overline{E_1} \cdot \operatorname{curl} E_1\}\, ds$$

$$+ 2i \operatorname{Im} \left\{ \int_{|y|=R} \{\nu \times E_1 \cdot \operatorname{curl} \overline{E_2} - \nu \times \overline{E_2} \cdot \operatorname{curl} E_1\}\, ds \right\}.$$

By using the series expansions for E_1 and E_2 we have

$$\int_{|y|=R} \{\nu \times E_1 \cdot \operatorname{curl} \overline{E_1} - \nu \times \overline{E_1} \cdot \operatorname{curl} E_1\}\, ds$$

$$= -\frac{2i}{\kappa} \sum_{l=1}^{\infty} \sum_{k=-l}^{l} |\alpha_l^k|^2 l(l+1) - 2i\kappa \sum_{l=1}^{\infty} \sum_{k=-l}^{l} |\beta_l^k|^2 l(l+1)$$

and

$$\int_{|y|=R} \{\nu \times E_1 \cdot \operatorname{curl} \overline{E_2} - \nu \times \overline{E_2} \cdot \operatorname{curl} E_1\}\, ds$$

$$= -\sum_{l=1}^{\infty} \sum_{k=-l}^{l} |\alpha_l^k|^2 l(l+1) \frac{4\pi}{\kappa^2 a^2 j_l(\kappa a)}$$

$$- \sum_{l=1}^{\infty} \sum_{k=-l}^{l} |\beta_l^k|^2 l(l+1) \frac{4\pi\kappa}{a(j_l(\kappa a) + \kappa a j_l'(\kappa a))} \in \mathbb{R}.$$

Analogous to Theorem 2 we can now conclude that $E_g(x) = 0$ in D_e. We now have that $g = 0$ by the L^2 jump relations for vector potentials [10], [11], and the proof is finished.

Now let $\tilde{E}$, $\tilde{H}\in C^1(\mathbb{R}^3\backslash\{0\})$ be a radiating solution to Maxwell's equations with electric far field pattern $\tilde{E}_\infty$. For $\varepsilon>0$ we can choose constants a_l^k, $b_l^k\in\mathbb{C}$, $l=1,\dots,N$, $k=-l,\dots,l$ such that

$$\left\| \nu\times\left(\tilde{E}+\sum_{l=1}^{N}4\pi i^l\frac{1}{[l(l+1)]^{1/2}}\sum_{k=-l}^{l}a_l^k\left(\operatorname{curl} M_l^k-\frac{i\kappa a^2 j_l(\kappa a)}{4\pi}\operatorname{curl} N_l^k\right)\right.\right.$$
$$\left.\left.-\sum_{l=1}^{N}4\pi i^l\frac{i\kappa}{[l(l+1)]^{1/2}}\sum_{k=-l}^{l}b_l^k\left(M_l^k-\frac{ia(j_l(\kappa a)+\kappa a j_l'(\kappa a))}{4\pi}N_l^k\right)\right)\right\|_{L^2(\partial D)}\leqq\varepsilon.$$

Define

$$g_N:=\sum_{l=1}^{N}\sum_{k=-l}^{l}a_l^kU_l^k+b_l^kV_l^k,$$

$$(4.6)\qquad \begin{aligned} E^i_{g_N}(x)&:=i\kappa\int_\Omega \exp(i\kappa x\cdot d)g_N(d)\,ds(d)\\ &=\sum_{l=1}^{N}4\pi i^l\frac{1}{[l(l+1)]^{1/2}}\sum_{k=-l}^{l}a_l^k\operatorname{curl} M_l^k-i\kappa b_l^kM_l^k, \qquad x\in\mathbb{R}^3,\end{aligned}$$

$$E^s_{g_N}(x):=\int_\Omega E^s(x,d,g_N(d))\,ds(d), \qquad x\in\bar{D}_e,$$

and

$$E^0_{g_N}(x):=a^2\sum_{l=1}^{N}i^l\frac{1}{[l(l+1)]^{1/2}}$$
$$\cdot\sum_{k=-l}^{l}\left(-a_l^ki\kappa j_l(\kappa a)\operatorname{curl} N_l^k(x)-b_l^k\frac{\kappa}{a}(j_l(\kappa a)+\kappa a j_l'(\kappa a))N_l^k(x)\right), \qquad |x|>0.$$

Then, for $|x|>a$,

$$E^0_{g_N}(x)=a^2\operatorname{curl}\int_\Omega\Phi(x,ad)(\mathbf{R}g_N)(d)\,ds(d),$$

and hence

$$\|\nu\times(\tilde{E}-(E^s_{g_N}-E^0_{g_N}))\|_{L^2(\partial D)}\leqq\varepsilon$$

and

$$\sup_{\hat{x}\in\Omega}|(\mathbf{F}g_N)(\hat{x})-\tilde{E}_\infty(\hat{x})|\leqq C\varepsilon$$

for some positive constant C. We can now conclude that if E^i_g and E^0_g are defined by (4.3), E_∞ is the exact far field data and S is the set of C^2 surfaces bounding a domain D containing the origin such that D_e is connected, then the infimum of the functional

$$M_3(g,\Gamma):=\|\mathbf{F}g-\tilde{E}_\infty\|_{L^2(\Omega)}+\|\nu\times(\tilde{E}+E^i_g+E^0_g)\|_{L^2(\Gamma)}$$

for $g\in T^2(\Omega)$ and $\Gamma\in S$ is zero for all $\kappa>0$, and the solution of the inverse electromagnetic obstacle problem is the limit of an appropriate minimizing sequence.

5. The inverse inhomogeneous medium problem for electromagentic waves. In this section we assume that $n \in C^{1,\alpha}(\mathbb{R}^3)$ is a real valued function such that $n(x) \neq 0$ for all $x \in \mathbb{R}^3$ and $m := 1 - n$ has compact support. Again let $a > 0$ be such that $j_l(\kappa a)(j_l(\kappa a) + \kappa a j_l'(\kappa a)) \neq 0$ for all $l \in \mathbb{N}$. Let B, $\tilde{B}$, $\hat{B}$ be open balls with center at the origin such that

$$\operatorname{supp}(m) \subset \hat{B} \subset B \subset \tilde{B}$$

and $B := \{x \in \mathbb{R}^3 \colon |x| < R\}$, where $R > a$. We again define the operator $\mathbf{F}$ by (4.2), where $E_\infty(\hat{x}, p, d)$ is the electric far field pattern corresponding to the incident field (1.1) being scattered by the inhomogeneous medium. We first prove injectivity of the operator $\mathbf{F}$.

THEOREM 7. *Assume* $\kappa > 0$. *If* $g \in T^2(\Omega)$ *is a solution of* $\mathbf{F}g = 0$, *then* $g = 0$.

Proof. The proof is analogous to the proof of Theorem 5, where we use Gauss' theorem and (1.7) to arrive at

$$\int_{|y|=R} \{\nu \times (E_g^i + E_g^s) \cdot \operatorname{curl} \overline{(E_g^i + E_g^s)} - \nu \times \overline{(E_g^i + E_g^s)} \cdot \operatorname{curl}(E_g^i + E_g^s)\}\, ds = 0.$$

We will now prove the analogue of Theorem 4. We define

$$X := \operatorname{span}\{4\pi \operatorname{curl} M_l^k - ia^2\kappa j_l(\kappa a) \operatorname{curl} N_l^k,$$

$$4\pi M_l^k - ia\{j_l(\kappa a) + \kappa a j_l'(\kappa a)\} N_l^k,\ l \in \mathbb{N},\ k = -l, \dots, l\},$$

$$Y := \{(F, G) \in (C^1(\bar{B}))^2 \colon \operatorname{curl} F - i\kappa G = 0, \operatorname{curl} G + i\kappa n F = 0 \text{ in} B\},$$

and

$$W := \{(\nu \times (E - F), \nu \times \operatorname{curl}(E - F)) \colon E \in X, (F, G) \in Y\} \subset T^2(\partial B) \times T^2(\partial B).$$

THEOREM 8. *W is dense in* $T^2(\partial B) \times T^2(\partial B)$.

Proof. Suppose a, $b \in T^2(\partial B)$ satisfy

$$\int_{\partial B} \{\bar{a} \cdot [\nu \times (E - F)] + \bar{b} \cdot [\nu \times \operatorname{curl}(E - F)]\}\, ds = 0 \tag{5.1}$$

for all $E \in X$, $(F, G) \in Y$. We must show that $a = b = 0$. To this end, we define

$$U^s(x) := \operatorname{curl} \int_{\partial B} \nu(y) \times \bar{b}(y) \Phi(x, y)\, ds(y)$$

$$+ \frac{1}{\kappa^2} \operatorname{curlcurl} \int_{\partial B} \nu(y) \times \bar{a}(y) \Phi(x, y)\, ds(y), \qquad x \in \mathbb{R}^3 \backslash \partial B,$$

$$U^0(x) := \frac{1}{\kappa^2} \operatorname{curl} \int_{\partial B} \chi(x, y)(\nu(y) \times \bar{a}(y))\, ds(y)$$

$$+ \frac{1}{\kappa^2} \operatorname{curl} \int_{\partial B} \tilde{\chi}(x, y)(\nu(y) \times \bar{b}(y))\, ds(y),\ x \in \mathbb{R}^3,$$

and $U(x) := U^0(x) + U^s(x)$, $x \in \mathbb{R}^3 \backslash \partial B$, where χ is defined by (4.5) and $\tilde{\chi}(x, y) := \operatorname{curl}_y \chi(x, y)$ (the curl is applied to the row vectors of χ).

We compute for $r < R$ that

$$\int_\Omega U(r\hat{x}) \cdot U_l^k(\hat{x})\, ds(\hat{x}) = \int_\Omega U(r\hat{x}) \cdot V_l^k(\hat{x})\, ds(\hat{x}) = 0$$

for all $l \in \mathbb{N}$, $k = -l, \dots, l$, and hence $U = 0$ in B.

Defining $U_+ := U|_{\mathbb{R}^3\setminus\bar{B}}$ we have by the L^2 jump relations for vector potentials (see [10], [11]) that $\nu\times U_+ = \nu\times\bar{b}$ and $\nu\times\operatorname{curl} U_+ = \nu\times\bar{a}$ on ∂B in a L^2 sense. Setting $E=0$ in (5.1) and inserting these equalities into (5.1) now shows that

$$0=\int_{\partial B}\{\nu\times\operatorname{curl} U_+\cdot F+\nu\times U_+\cdot\operatorname{curl} F\}\,ds \tag{5.2}$$

for all $(F, G)\in Y$.

We now compute

$$\begin{aligned}U^0(x)&=4\pi\sum_{l=1}^{\infty}\sum_{k=-l}^{l} i^l\frac{1}{[l(l+1)]^{1/2}}\alpha_l^k\operatorname{curl} M_l^k(x)\\&\quad-4\pi\sum_{l=1}^{\infty}\sum_{k=-l}^{l} i\kappa i^l\frac{1}{[l(l+1)]^{1/2}}\beta_l^k M_l^k(x),\qquad x\in\mathbb{R}^3,\\U^s(x)&=-i\kappa a^2\sum_{l=1}^{\infty}\sum_{k=-l}^{l} i^l\frac{j_l(\kappa a)}{[l(l+1)]^{1/2}}\alpha_l^k\operatorname{curl} N_l^k(x)\\&\quad-\kappa a\sum_{l=1}^{\infty}\sum_{k=-l}^{l} i^l\frac{j_l(\kappa a)+\kappa a j_l'(\kappa a)}{[l(l+1)]^{1/2}}\beta_l^k N_l^k(x)\end{aligned}$$

for $|x|>R$, where

$$\begin{aligned}\alpha_l^k&=-\frac{(-i)^l}{[l(l+1)]^{1/2}a^2 j_l(\kappa a)}\left\{\frac{1}{\kappa^2}\int_{\partial B}\operatorname{curl}\overline{M_l^k(y)}\cdot(\nu(y)\times\overline{a(y)})\,ds(y)\right.\\&\qquad\left.+\int_{\partial B}\overline{M_l^k(y)}\cdot(\nu(y)\times\overline{b(y)})\,ds(y)\right\},\\\beta_l^k&=-\frac{(-i)^{(l+1)}}{[l(l+1)]^{1/2}a(j_l(\kappa a)+\kappa a j_l'(\kappa a))}\left\{\int_{\partial B}\overline{M_l^k(y)}\cdot(\nu(y)\times\overline{a(y)})\,ds(y)\right.\\&\qquad\left.+\int_{\partial B}\operatorname{curl}\overline{M_l^k(y)}\cdot(\nu(y)\times\overline{b(y)})\,ds(y)\right\}.\end{aligned}$$

If we now set

$$g_N:=\sum_{l=1}^{N}\sum_{k=-l}^{l}\alpha_l^k U_l^k+\beta_l^k V_l^k,$$

$$\begin{aligned}U_N^0(x)&:=\int_\Omega E^i(x,d,g_N(d))\,ds(d)\\&=i\kappa\int_\Omega \exp(i\kappa x\cdot d)g_N(d)\,ds(d),\qquad x\in\mathbb{R}^3,\end{aligned}\tag{5.3}$$

$$E_{g_N}^s(x)=\int_\Omega E^s(x,d,g_N(d))\,ds(d),\qquad x\in\mathbb{R}^3,$$

then from (4.6) U_N^0 converges uniformly together with its derivatives on compact subsets of $\mathbb{R}^3$ to U^0 as $N\to\infty$, and $E_{g_N}^s$ converges uniformly together with its derivatives on $\mathbb{R}^3$ to E_g^s as $N\to\infty$, where E_g^s denotes the scattered electric field corresponding to the incident field U^0.

We will now show that E_g^s and U^s coincide outside of a large sphere. We first note [3] that the vector Lippmann-Schwinger equation can be written as

$$E(x)=E^i(x)-\operatorname{curlcurl}\int_{\mathbb{R}^3}\Phi(x,y)m(y)E(y)\,dy,\qquad x\in\mathbb{R}^3\setminus\bar{B},$$

i.e.,

$$E^s(x) = -\text{curlcurl} \int_{\mathbb{R}^3} \Phi(x, y) m(y) E(y)\, dy, \qquad x \in \mathbb{R}^3 \backslash \bar{B},$$

is the scattered electric field corresponding to the incident field E^i. From this we deduce that for $\hat{x}$, $d \in \Omega$, $p \in \mathbb{R}^3$, the electric far field pattern E_∞ can be represented in terms of the total field E by

$$(5.4) \qquad E_\infty(\hat{x}, d, p) = -\frac{\kappa^2}{4\pi} \int_B \exp(-i\kappa \hat{x} \cdot y) m(y) (\hat{x} \times E(y, d, p)) \times \hat{x}\, dy.$$

Using the far field representation for the far field pattern U_∞ of U^s obtained from the representation theorem for U^s and (5.2), we compute, for arbitrary $p \in \mathbb{R}^3$, $\hat{x} \in \Omega$, that

$$\begin{aligned}
&p \cdot U_\infty(\hat{x}) \\
&= \frac{i\kappa}{4\pi} p \cdot \left\{ \hat{x} \times \int_{\partial B} \left(\nu(y) \times U^s(y) + \frac{1}{i\kappa} [\nu(y) \times \text{curl } U^s(y)] \times \hat{x} \right) \exp(-i\kappa \hat{x} \cdot y)\, ds(y) \right\} \\
&= \frac{1}{4\pi} \int_{\partial B} \Big(\nu(y) \times U^s(y) \cdot \text{curl}_y \{p \exp(-i\kappa \hat{x} \cdot y)\} \\
(5.5) \qquad &\qquad - \nu(y) \times \frac{i}{\kappa} \text{curl}_y\, \text{curl}_y \{p \exp(-i\kappa \hat{x} \cdot y)\} \cdot \frac{1}{i\kappa} \text{curl } U^s(y) \Big)\, ds(y) \\
&= \frac{1}{4\pi} \int_{\partial B} \left(\nu(y) \times U^s(y) \cdot H(y, -\hat{x}, p) - \nu(y) \times E(y, -\hat{x}; p) \cdot \frac{1}{i\kappa} \text{curl } U^s(y) \right) ds(y) \\
&= -\frac{1}{4\pi} \int_{\partial B} \left(\nu(y) \times U^0(y) \cdot H(y, -\hat{x}, p) - \nu(y) \times E(y, -\hat{x}, p) \cdot \frac{1}{i\kappa} \text{curl } U^0(y) \right) ds(y) \\
&= \frac{i\kappa}{4\pi} \int_B m(y) U^0(y) \cdot E(y, -\hat{x}, p)\, dy.
\end{aligned}$$

Hence, from (5.4) and (5.5) we arrive at

$$\begin{aligned}
p \cdot E_{g,\infty}(\hat{x}) &= \lim_{N \to \infty} p \cdot E_{g_N,\infty}(\hat{x}) \\
&= \lim_{N \to \infty} \int_\Omega p \cdot E_\infty(\hat{x}, d, g_N(d))\, ds(d) \\
&= \lim_{N \to \infty} \int_\Omega g_N(d) \cdot E_\infty(-d, -\hat{x}, p)\, ds(d) \\
&= \lim_{N \to \infty} \int_\Omega g_N(d) \\
&\qquad \cdot \int_B -\frac{\kappa^2}{4\pi} \exp(i\kappa d \cdot y) m(y) \{(d \times E(y, -\hat{x}, p)) \times d\}\, dy\, ds(d) \\
&= \lim_{N \to \infty} \int_B -\frac{\kappa^2}{4\pi} \frac{1}{i\kappa} \int_\Omega i\kappa g_N(d) \\
&\qquad \cdot \exp(i\kappa d \cdot y) m(y) \{(d \times E(y, -\hat{x}, p)) \times d\}\, ds(d)\, dy \\
&= \frac{i\kappa}{4\pi} \int_B m(y) E(y, -\hat{x}, p) \cdot U^0(y)\, dy \\
&= p \cdot U_\infty(\hat{x}),
\end{aligned}$$

i.e., $U_\infty(\hat{x}) = E_{g,\infty}(\hat{x})$ for all $\hat{x} \in \Omega$. We can now conclude by Rellich's lemma [2], [3], that $U^s(x) = E_g^s(x)$ for $|x|$ sufficiently large.

To complete the proof we will now show that $U = U^s + U^0$ vanishes identically, and hence by the L^2 jump relations for vector potentials $a = b = 0$. By Green's theorem we have for $R_1 > R$,

$$\begin{aligned} 0 &= \int_{|y|=R_1} \{\nu \times (U^0 + E_g^s) \cdot \operatorname{curl} \overline{(U^0 + E_g^s)} \\ &\qquad - \nu \times \overline{(U^0 + E_g^s)} \cdot \operatorname{curl} (U^0 + E_g^s)\}\, ds \\ &= \int_{|y|=R_1} \{\nu \times (U^0 + U^s) \cdot \operatorname{curl} \overline{(U^0 + U^s)} \\ &\qquad - \nu \times \overline{(U^0 + U^s)} \cdot \operatorname{curl} (U^0 + U^s)\}\, ds \\ &= \int_{|y|=R_1} \{\nu \times U^s \cdot \operatorname{curl} \overline{U^s} - \nu \times \overline{U^s} \cdot \operatorname{curl} U^s\}\, ds \\ &\quad + 2i \operatorname{Im}\left\{\int_{|y|=R_1} \{\nu \times U^s \cdot \operatorname{curl} \overline{U^0} - \nu \times \overline{U^0} \cdot \operatorname{curl} U^s\}\, ds\right\} \end{aligned} \tag{5.6}$$

and

$$\begin{aligned} &\int_{|y|=R_1} \{\nu \times U^s \cdot \operatorname{curl} \overline{U^s} - \nu \times \overline{U^s} \cdot \operatorname{curl} U^s\}\, ds \\ &\qquad = -2i\kappa^3 a^4 \sum_{l=1}^{\infty} \sum_{k=-l}^{l} j_l(\kappa a)^2 |\alpha_l^k|^2 \\ &\qquad\quad - 2i\kappa a^2 \sum_{l=1}^{\infty} \sum_{k=-l}^{l} (j_l(\kappa a) + \kappa a j_l'(\kappa a))^2 |\beta_l^k|^2, \end{aligned} \tag{5.7}$$

$$\begin{aligned} &\int_{|y|=R_1} \{\nu \times U^s \cdot \operatorname{curl} \overline{U^0} - \nu \times \overline{U^0} \cdot \operatorname{curl} U^s\}\, ds \\ &\qquad = -\sum_{l=1}^{\infty} \sum_{k=-l}^{l} \kappa^2 a^2 j_l(\kappa a) 4\pi |\alpha_l^k|^2 \\ &\qquad\quad + \sum_{l=1}^{\infty} \sum_{k=-l}^{l} 4\pi\kappa a (j_l(\kappa a) + \kappa a j_l'(\kappa a)) |\beta_l^k|^2 \in \mathbb{R}. \end{aligned} \tag{5.8}$$

Inserting (5.7) and (5.8) into (5.6) we can now conclude that $\alpha_l^k = \beta_l^k = 0$ for all $l \in \mathbb{N}$, $k = -l, \ldots, l$, i.e., $U(x) = 0$ for $x \in \mathbb{R}^3 \backslash \bar{B}$, and hence $a = b = 0$ by the L^2 jump relations for vector potentials [10], [11]. This finishes the proof.

For $\tilde{E}, \tilde{H} \in C^1(\mathbb{R}^3 \backslash \{0\})$, a radiating solution to Maxwell's equations with electric far field pattern $\tilde{E}_\infty$, we now define the cost functional

$$\begin{aligned} M_4(g, F, G, m) := \|\mathrm{F}g - \tilde{E}_\infty\|_{L^2(\Omega)} + \|\hat{x} \times (F - \tilde{E} - E_g^i - E_g^0)\|_{L^2(\partial B)} \\ + \|\hat{x} \times \operatorname{curl} (F - \tilde{E} - E_g^i - E_g^0)\|_{L^2(\partial B)} \end{aligned}$$

for $m \in C_0^{1,\alpha}(\hat{B})$, $(F, G) \in Y$ and E_g^i, E_g^0 defined by (4.3). To prove that the infimum of the functional M_4 is zero for exact electric far field data E_∞ and all $\kappa > 0$ define

$m := 1-n$ and

$$g_N := \sum_{l=1}^{N} \sum_{k=-l}^{l} \alpha_l^k U_l^k + \beta_l^k V_l^k,$$

$$E_{g_N}^i(x) := i\kappa \int_\Omega \exp(i\kappa x\cdot d) g_N(d)\, ds(d), \qquad x \in \mathbb{R}^3,$$

$$E_{g_N}^s(x) := \int_\Omega E^s(x, d, g_N(d))\, ds(d), \qquad x \in \mathbb{R}^3,$$

$$U_N^s(x) := \sum_{l=1}^{N} \sum_{k=-l}^{l} \alpha_l^k i^l \frac{1}{[l(l+1)]^{1/2}} (-ia^2\kappa j_l(\kappa a)) \operatorname{curl} N_l^k(x)$$

$$+ \sum_{l=1}^{N} \sum_{k=-l}^{l} \beta_l^k i^l \frac{i\kappa}{[l(l+1)]^{1/2}} ia(j_l(\kappa a) + \kappa a j_l'(\kappa a)) N_l^k(x), \qquad x \in \mathbb{R}^3\backslash\{0\},$$

$$U_N^s(x) = a^2 \operatorname{curl} \int_\Omega (\mathbf{R}g_N)(d)\Phi(x, ad)\, ds(d), \qquad |x| > a.$$

Then, from (4.6) and Theorem 8, for a given $\varepsilon > 0$ we can find α_l^k, $\beta_l^k \in \mathbb{C}$, $l = 1, \dots, N$, $k = -l, \dots, l$, and $(F, G) \in Y$ such that

$$\|\hat{x} \times (F - \tilde{E} - E_{g_N}^i - U_N^s)\|_{L^2(\partial B)} + \|\hat{x} \times \operatorname{curl}(F - \tilde{E} - E_{g_N}^i - U_N^s)\|_{L^2(\partial B)} \leqq \varepsilon.$$

Defining

$$F^i(x) := -\operatorname{curl} \int_{\partial B} \nu(y) \times F(y)\Phi(x, y)\, ds(y) + \operatorname{grad} \int_{\partial B} \nu(y) \cdot F(y)\Phi(x, y)\, ds(y)$$

$$- \int_{\partial B} \nu(y) \times \operatorname{curl} F(y)\Phi(x, y)\, ds(y), \qquad x \in B,$$

$$F^s(x) := \operatorname{grad} \int_B \frac{1}{n} \operatorname{grad} n(y) F(y)\Phi(x, y)\, dy - \kappa^2 \int_B m(y)F(y)\Phi(x, y)\, dy, \qquad x \in \mathbb{R}^3,$$

we have from the Stratton-Chu formula [3, Thm. 6.1] that $F(x) = F^i(x) + F^s(x)$ in B. Since $\operatorname{div} F^i = 0$ in B $(F^i, (1/i\kappa) \operatorname{curl} F^i)$ is a solution to the Maxwell equations in B. On the other hand,

$$F^s(x) = -\operatorname{curlcurl} \int_B m(y)F(y)\Phi(x, y)\, dy, \qquad x \in \mathbb{R}^3\backslash\bar{B},$$

is the scattered part of the inhomogeneous medium problem solution corresponding to the incident field F^i. Using the representation theorems [2], [3], we now arrive at

$$\|F^i - E_{g_N}^i\|_{\infty,\tilde{B}} \leqq C_1\varepsilon, \tag{5.9}$$

$$\|\hat{x} \times (F^s - \tilde{E} - U_N^s)\|_{L^2(\partial\tilde{B})} \leqq C_2\varepsilon \tag{5.10}$$

for positive constants C_1, C_2. We, therefore, conclude from (5.9) and the vector Lippmann-Schwinger equation that

$$\|F^s - E_{g_N}^s\|_{\infty,\partial\tilde{B}} \leqq C_3\varepsilon, \tag{5.11}$$

and from (5.10) and (5.11),

$$\|E_{g_N,\infty} - \tilde{E}_\infty - U_{N,\infty}\|_{\infty,\Omega} \leqq C_4\varepsilon,$$

i.e.,

$$\|\mathbf{F}g_N - \tilde{E}_\infty\|_{\infty,\Omega} \leqq C_4\varepsilon$$

for positive constants C_3, C_4. It now follows that the infimum of M_4 is zero for all $\kappa > 0$, and the solution of the inverse electromagnetic inhomogeneous medium problem is the limit of an appropriate minimizing sequence.

REFERENCES

[1] J. BLÖHBAUM, *Optimisation methods for an inverse problem with time-harmonic electromagnetic waves: an inverse problem in electromagnetic scattering*, Inverse Probl., 5 (1989), pp. 463–482.

[2] D. COLTON AND R. KRESS, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, 1983.

[3] ———, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, Berlin, 1992.

[4] ———, *Time harmonic electromagnetic waves in an inhomogeneous medium*, Proc. Roy. Soc. Edinburgh, 116A (1990), pp. 279–293.

[5] D. COLTON AND P. MONK, *The numerical solution of the three dimensional inverse scattering problem for time harmonic acoustic waves*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 278–291.

[6] ———, *The inverse scattering problem for acoustic waves in an inhomogeneous medium*, Quart. J. Mech. Appl. Math., 41 (1988), pp. 97–125.

[7] ———, *A new method for solving the inverse scattering problem for acoustic waves in an inhomogeneous medium*, Inverse Prob., 5 (1989), pp. 1013–1026.

[8] ———, *On a class of integral equations of the first kind in inverse scattering theory*, SIAM J. Appl. Math., to appear.

[9] 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.

[10] P. HÄHNER, *An exterior boundary-value problem for the Maxwell equations with boundary data in a Sobolev space*, Proc. Roy. Soc. Edinburgh, 109A (1988), pp. 213–224.

[11] ———, *An approximation theorem in inverse electromagnetic scattering*, Math. Methods Appl. Sci., to appear.

[12] H. KERSTEN, *Grenz- und Sprungrelationen für Potentiale mit quadratsummierbarer Dichte*, Resultate Math., 3 (1980), pp. 17–24.

SIAM J. MATH. ANAL.
Vol. 24, No. 2, pp. 390-406, March 1993

DECOMPOSITION OF THE DISPLACEMENT VECTOR FIELD AND DECAY RATES IN LINEAR THERMOELASTICITY*

JAIME E. MUÑOZ RIVERA†

Abstract. The linear thermoelastic system is studied in the whole space $\mathbb{R}^n$, and it is proved that the displacement vector field can be decomposed into two parts, one that conserves its energy, and the other that decays uniformly to zero as time goes to infinity. The method used here is based on a new Lyapunov function.

Key words. linear thermoelasticity, energy decay rates

AMS(MOS) subject classifications. 35B40, 35M05

1. Introduction. In this work we will study the classical linear thermoelastic system for homogeneous isotropic materials. The system in question consists of n hyperbolic equations of motion coupled with the parabolic equation of energy. A first work about the asymptotic behavior of the solution of the thermoelastic system is given by Dafermos [3], who proved that they are asymptotically stable when time goes to infinity; in that work regularity and uniqueness results are also proved. Here we will study the decomposition of the displacement vector field and the asymptotic behaviour of the energy associated with each part of this decomposition. The question of partition of the energy was first investigated by Lax and Phillips [7] for a classical solution of the wave equation. There the authors established that the kinetic and the potential energies become asymptotically equal as time goes to infinity. Another work in this direction is that of Duffin [5], who proved that the equipartition is consummated in a finite time provided the initial data have compact support. Corresponding results for abstract hyperbolic equations have been obtained by Goldstein [6], Costa and Strauss [2], among others. It seems that a first work about the partition of the energy in the framework of isotropic linear thermoelasticity (system with dissipation) was given by Dassios and Grillakis [4]. There the authors studied how the energy associated with the longitudinal and thermal wave is divided into kinetic, strain, and thermal energy in the case $\Omega=\mathbb{R}^3$. They concluded that the rate of decay of the energy stored in the longitudinal and thermal wave is affected by the symmetric of the initial data, that is, if v_0, v_1, v_θ are the lowest nonvanishing moments of the initial displacement, initial velocity and initial temperature, respectively, then all of the three parts of the energy associated with the longitudinal and thermal wave decay to zero as $t\rightarrow+\infty$ at the rate $t^{-(m+3/2)}$, where $m=\min\{v_0+1, v_1, v_\theta\}$, and whenever the initial data have continuous fifth derivatives with compact support in $\mathbb{R}^3$, while the transverse wave conserves its energy.

For bounded anisotropic and inhomogeneous bodies Chirita [1] proved that the mean thermal energy tends to zero as time goes to infinity and that the asymptotic equipartition occurs between the Cesàro means of the kinetic and strain energies, which implies that the thermal effects do not influence explicitly the asymptotic equipartition of the mean kinetic and strain energies.

In special situations, that is, when the restoring force is proportional to the vector velocity of the displacement vector field, Pereira and Menzala [10] proved that in a

* Received by the editors June 12, 1991; accepted for publication June 1, 1992.

† National Laboratory for Scientific Computation (LNCC-CNPQ), Rua Lauro Muller 455, Botafogo, CEP. 22290, Rio de Janeiro, RJ, Brazil, and IMUFR, P.O. Box 68530, Rio de Janeiro, RJ, Brazil.

bounded, isotropic and inhomogeneous body the kinetic, strain, and thermal energy approach zero exponentially as $t \to +\infty$.

In this work we will study the decomposition of the displacement vector field in $\mathbb{R}^n$ $(n>1)$ into two parts. One of them, the solenoidal part, that is, the nondissipative component that conserves its energy and the other, the dissipative component (irrotational) part for $n=3$), decays to zero when t approaches infinity. We will prove (§ 3) that both the displacement vector field in the dissipative direction and the deviation of the temperature decay as $t^{-n/2}$ when t tends to infinity. For $n=1$, the system is totally dissipative in the sense that the total energy decays to zero as $t \to 0$ as fast as $t^{-1/2}$. Moreover, we will find a parameter m for which the rate of decay increases with increasing values of m. More precisely we shall prove that the rate of decay of the energy increases as $t^{-(m+n/2)}$ if each component of the initial data is the m-derivative of a function that belongs to $L^2(\mathbb{R}^n) \cap L^1(\mathbb{R}^n)$.

Through this result we improve the work of Dassios and Grillakis [4], which proved that such a decomposition exists for the displacement vector field in $\mathbb{R}^3$. The method we use here is different and simpler than that used in [4].

The outline of this paper is as follows: In § 2 we consider the Cauchy problem associated to the thermoelastic system. The existence of global solution under more general assumptions is shown in [3], but for the sake of completeness we briefly indicate the proof in the framework of semigroups of operators. In § 3 we study the asymptotic properties of the density function for the thermoelastic system with initial data in the dissipative direction by using a method based on the construction of a suitable Lyapunov's functional. Finally, in § 4 we study the decomposition of the displacement vector field when the initial data is taken in the domain of the operator associated with elliptic part of the thermoelastic system.

2. Existence, uniqueness, and regularity. In the absence of body forces and heat sources the linear thermoelastic system of a homogeneous and isotropic body occupies all space $\mathbb{R}^n$ is given by

$$\mathbf{u}_{tt} - (a^2 - b^2)\Delta \mathbf{u} - b^2 \nabla\{\operatorname{div} \mathbf{u}\} + \alpha \nabla \theta = 0 \quad \text{in } \mathbb{R}^n_x \times \mathbb{R}_t, \tag{2.1}$$

$$\theta_t - k\Delta\theta + \beta \operatorname{div} \mathbf{u}_t = 0 \quad \text{in } \mathbb{R}^n_x \times \mathbb{R}_t, \tag{2.2}$$

$$\mathbf{u}(x,0) = \mathbf{u}_0(x), \mathbf{u}_t(x,0) = \mathbf{u}_1(x), \theta(x,0) = \theta_0(x) \quad \text{in } \mathbb{R}^n_x, \tag{2.3}$$

where $a^2 = (\mu+\lambda)/\rho$, $b^2 = \lambda/\rho$, $\alpha = \gamma/\rho$, $\beta = \nu/k$. We are denoting by μ and λ the Lamé constants; by ρ the mass density, k is defined by $k = \lambda_0/c_\varepsilon$, where λ_0 is the thermal conductivity and c_ε is the specific heat. Finally ν is the thermal diffusivity.

Let's denote by $\mathbf{A}$ the operator on $[L^2(\mathbb{R}^n)]^n$ and by A the operator on $L^2(\mathbb{R}^n)$ with domain $D(\mathbf{A}) = [H^2(\mathbb{R}^n)]^n$, $D(A) = H^2(\mathbb{R}^n)$, respectively, such that

$$\mathbf{A}w = -(a^2 - b^2)\Delta w - b^2 \nabla \operatorname{div} \mathbf{w},$$

$$Aw = -k\Delta w,$$

where Δ, ∇, div stands for the Laplacian, the gradient and the divergence operator, $\mathbf{w}$ is a vector $\mathbf{w} = (w_1, \dots, w_n)$, and for $\Delta\mathbf{w}$ we are denoting $\Delta\mathbf{w} = (\Delta w_1, \dots, \Delta w_n)$.

It is well known that $\mathbf{A}$ and A are positive selfadjoint operators in the Hilbert space $[L^2(\mathbb{R}^n)]^n$ and $L^2(\mathbb{R}^n)$, respectively.

Let us define the space $\mathcal{H}$ as

$$\mathcal{H} = [H^1(\mathbb{R}^n)]^n \times [L^2(\mathbb{R}^n)]^n \times L^2(\mathbb{R}^n)$$

with norm given by

$$\|V\|_{\mathscr{H}}^2=\int_{\mathbb{R}^n}|\mathbf{u}|^2\,d\mathbf{x}+(a^2-b^2)\sum_{i=1}^{n}\int_{\mathbb{R}^n}|\nabla u_i|^2\,d\mathbf{x}+b^2\int_{\mathbb{R}^n}|\operatorname{div} u|^2\,d\mathbf{x}$$
$$+\int_{\mathbb{R}^n}|\mathbf{v}|^2\,d\mathbf{x}+\int_{\mathbb{R}^n}|\theta|^2\,d\mathbf{x}$$

for any $V=(\mathbf{u},\mathbf{v},\theta)^\tau$, where $\mathbf{u}=(u_1,\ldots,u_n)$ and $\mathbf{v}=(v_1,\ldots,v_n)$. We introduce the operator

$$\mathscr{A}\colon D(\mathscr{A})\subseteq\mathscr{H}\to\mathscr{H}$$
$$\mathscr{A}=-\begin{pmatrix}0 & -I & 0\\ \mathbf{A} & 0 & \alpha B\\ 0 & \beta B^* & A\end{pmatrix}$$

with domain

$$D(\mathscr{A})=[H^2(\mathbb{R}^n)]^n\times[H^1(\mathbb{R}^n)]^n\times H^2(\mathbb{R}^n),$$

where B is defined as $Bw=\nabla w$ with domain $D(B)=H^1(\mathbb{R}^n)$.

The existence result is stated in the following theorem.

THEOREM 2.1 (Existence and uniqueness). *Let* $(\mathbf{u}_0,\mathbf{u}_1,\theta_0)^\tau\in D(\mathscr{A})$ *and* $T>0$, *then for any* $\alpha,\beta\in\mathbb{R}$, $k>0$, *there exist only one strong solution of system* (2.1)-(2.3) *satisfying*

$$\mathbf{u}\in\mathbf{C}(0,T;[H^2(\mathbb{R}^n)]^n)\cap\mathbf{C}^1(0,T;[H^1(\mathbb{R}^n)]^n)\cap\mathbf{C}^2(0,T;[L^2(\mathbb{R}^n)]^n)$$
$$\theta\in\mathbf{C}(0,T;H^2(\mathbb{R}^n))\cap\mathbf{C}^1(0,T;L^2(\mathbb{R}^n)).$$

Proof. System (2.1)-(2.3) is equivalent to

$$\frac{d}{dt}U=\mathscr{A}U,\quad U(0)=U_0,\quad U\in D(\mathscr{A}),$$

where

$$U=\begin{pmatrix}\mathbf{u}\\ \mathbf{u}_t\\ \theta\end{pmatrix},\quad U_0=\begin{pmatrix}\mathbf{u}_0\\ \mathbf{u}_1\\ \theta_0\end{pmatrix}.$$

In order to establish the existence result it is sufficient to show that $\mathscr{A}$ is the infinitesimal generator of a strongly continuous semigroup. To prove this, let's define the operator $\mathscr{B}=-[1+(\alpha-\beta)^2]I+\mathscr{A}$, with $D(\mathscr{B})=D(\mathscr{A})$, where I denotes the identity operator in $\mathscr{H}$. First we prove that $\mathscr{B}$ is dissipative. Let us take $V=(\mathbf{u},v,\theta)^\tau$ in $D(\mathscr{A})$; then it follows that

$$(\mathscr{A}V,V)_{\mathscr{H}}=-\int_{\mathbb{R}^n}\mathbf{u}\cdot\mathbf{v}\,d\mathbf{x}+(\alpha-\beta)\int_{\mathbb{R}^n}\mathbf{v}\cdot\nabla\theta\,d\mathbf{x}-k\int_{\mathbb{R}^n}|\nabla\theta|^2\,d\mathbf{x}$$
$$\leqq\frac{1}{2}[1+(\alpha-\beta)^2]\int_{\mathbb{R}^n}|\mathbf{v}|^2\,d\mathbf{x}+\frac{1}{2}\int_{\mathbb{R}^n}|\mathbf{u}|\,d\mathbf{x}-\frac{k}{2}\int_{\mathbb{R}^n}|\nabla\theta|^2\,d\mathbf{x}$$
$$\leqq\frac{1}{2}[1+(\alpha-\beta)^2]\|V\|_{\mathscr{H}}^2-\frac{k}{2}\int_{\mathbb{R}^n}|\nabla\theta|^2\,d\mathbf{x}.$$

Consequently, $\mathscr{B}$ satisfies

$$(\mathscr{B}V,V)_{\mathscr{H}}\leqq-\frac{1}{2}[1+(\alpha-\beta)^2]\|V\|_{\mathscr{H}}^2-\frac{k}{2}\int_{\mathbb{R}^n}|\nabla\theta|^2\,d\mathbf{x}.$$

Our next goal is to prove that $\operatorname{Im}[I-\mathscr{B}]=\mathscr{H}$, that is, for all $\mathbf{F}=(F_1, F_2, F_3)\in\mathscr{H}$ we can find $V=(u, v, \theta)\in D(\mathscr{A})=D(I-\mathscr{B})$ satisfying

$$\mu I+\mathscr{A}V=\mathbf{F};$$

here $\mu=2+(\alpha+\beta)^2$. In fact, by taking the Fourier transform to system and solving the result system we conclude that $\mathscr{B}$ is maximal monotone. Finally, since $D(\mathscr{B})$ is dense in $\mathscr{H}$, using Lummer-Phillips's theorem we conclude that $\mathscr{B}$ is the infinitesimal generator of a strongly continuous semigroup; so is $\mathscr{A}$. The required conclusion then follows. □

In order to obtain the decay of the energy associated to system (2.1)-(2.3), when the initial data is taken in $D(\mathscr{A})$, we will use the regularity result of system (2.1)-(2.3) and the density of $D(\mathscr{A}^3)$ in $D(\mathscr{A})$. The regularity and the density property are well known, but to assist the reader we briefly indicate the following remarks.

Remark 2.1. Let us define $D(\mathscr{A}^2)$ and $D(\mathscr{A}^3)$ as

$$D(\mathscr{A}^2)=\{V\in D(\mathscr{A});\ \mathscr{A}V\in D(\mathscr{A})\},$$

$$D(\mathscr{A}^3)=\{V\in D(\mathscr{A}^2);\ \mathscr{A}V\in D(\mathscr{A}^2)\}.$$

It is easy to see that whenever $U_0=(u_0, u_1, \theta_0)^\tau\in D(\mathscr{A}^3)$, we have

$$U=(u, u_t, \theta)^\tau\in\mathbf{C}^2(0, T; D(\mathscr{A}))$$

or, in particular,

$$u\in\mathbf{C}^3(0, T; H^2(\mathbb{R}^n)), \qquad \theta\in\mathbf{C}^2(0, T; H^2(\mathbb{R}^n)).$$

3. Asymptotic behaviour. The total energy in linear thermoelasticity for $\mathbb{R}^n$ in general does not decay to zero. In this section we will prove that the energy associated with the thermoelastic system decays to zero at the rate of $t^{-n/2}$ when time goes to infinity, provided the initial data is in the dissipative direction. Moreover, we will prove that the rate of decay increases if each component of the initial data is the m-derivative of a function that belongs to $L^2(\mathbb{R}^n)\cap L^1(\mathbb{R}^n)$; when this is the case we will have that the total energy decays as $t^{-(m+n/2)}$ when $t\to+\infty$.

Let $\alpha=(\alpha_1,\dots,\alpha_n)\in\mathbb{N}^n$ and $x=(x_1,\dots,x_n)\in\mathbb{R}^n$; then we shall denote by $|\alpha|$, $\mathbf{x}^\alpha$, and ∂^α the following expressions:

$$|\alpha|=\sum_{i=1}^n\alpha_i,$$

$$\mathbf{x}^\alpha=x_1^{\alpha_1}x_2^{\alpha_2}\cdots x_n^{\alpha_n},$$

$$\partial^\alpha=\left(\frac{\partial}{\partial x_1}\right)^{\alpha_1}\left(\frac{\partial}{\partial x_2}\right)^{\alpha_2}\cdots\left(\frac{\partial}{\partial x_n}\right)^{\alpha_n}.$$

Let us consider the following lemmas.

LEMMA 3.1. *Let v be a function in $L^2(\mathbb{R}^n)\cap L^1(\mathbb{R}^n)$ for which there exist $f\in L^2(\mathbb{R}^n)\cap L^1(\mathbb{R}^n)$ satisfying*

$$v=\partial^\alpha f, \quad \textit{where } |\alpha|=m.$$

Then we have

$$|\hat{v}(\xi)|\leqq[2\pi]^{-n/2}|\xi^\alpha|\int_{\mathbb{R}^n}|f(\mathbf{x})|\,d\mathbf{x}\quad\forall\xi\in\mathbb{R}^n, \tag{3.1}$$

where $\hat{v}$ denotes the usual Fourier transform of v.

Proof. Integration by parts of $\hat{v}$ implies that

$$\hat{v}(\xi)=[2\pi]^{-n/2}(-i)^{|\alpha|}|\xi^{\alpha}|\int_{\mathbb{R}^n} f(x)\, e^{ix\xi}\, dx, \tag{3.2}$$

where $i=\sqrt{-1}$. Since $f\in L^2(\mathbb{R}^n)\cap L^1(\mathbb{R}^n)$, (3.2) yields (3.1). □

LEMMA 3.2. *If r is a positive real number, then we have*

$$\int_0^1 \eta^{2m}\, e^{-r\eta^2}\, d\eta \leqq (2m)!\sqrt{n/2}\,r^{-m-1/2}, \quad m=0,1,\dots. \tag{3.3}$$

Proof. Let us denote by I_m the integral in the left-hand side of (3.3). Straightforward calculations give the inequality

$$I_0=\int_0^1 e^{-r\eta^2}\, d\eta \leqq \sqrt{n/2}\,r^{-1/2}. \tag{3.4}$$

Integration by parts of I_m with $m>0$ implies

$$I_m \leqq (2r)^{-1}(2m-1)\int_0^1 \eta^{2m-2}\, e^{-r\eta^2}\, d\eta \leqq (2r)^{-1}(2m-1)I_{m-1}.$$

Consequently,

$$I_m \leqq r^{-m}(2m)!I_0.$$

Substitution of (3.4) into the last inequality gives (3.3). □

LEMMA 3.3. *If the initial data $\mathbf{u}_0$ and $\mathbf{u}_1$ are such that*

$$D_iu_o^j(\mathbf{x})=D_ju_o^i(\mathbf{x}), \qquad D_iu_1^j(\mathbf{x})=D_ju_1^i(\mathbf{x}) \quad \textit{for } i\neq j,$$

where $\mathbf{u}_0=(u_0^1,\dots,u_0^n)$, $\mathbf{u}_1=(u_1^1,\dots,u_1^n)$ then the solution $\mathbf{u}$ of system (2.1)-(2.3) also satisfies

$$D_iu^j(\mathbf{x},t)=D_ju^i(\mathbf{x},t), \qquad D_iu_t^j(\mathbf{x},t)=D_ju_t^i(\mathbf{x},t) \quad \textit{in } \mathbb{R}_x^n\times\mathbb{R}_t.$$

Proof. From system (2.1)-(2.3), it is easy to see that

$$C_{ij}\mathbf{u}=D_iu^j-D_ju^i$$

satisfies

$$\{C_{ij}\mathbf{u}\}_{tt}-\Delta\{C_{ij}\mathbf{u}\}=0,$$

$$C_{ij}\mathbf{u}(x,0)=0, \qquad C_{ij}\mathbf{u}_t(x,0)=0.$$

By the uniqueness of the solution the result follows. □

LEMMA 3.4. *Suppose that*

$$D_iF_j(\mathbf{x},t)=D_jF_i(\mathbf{x},t)$$

holds for all $i\neq j$. Then we have

$$\left|\sum_{i=1}^n \xi_i\hat{F}_i\right|^2=|\xi|^2|\hat{F}|^2,$$

where $\xi=(\xi_1,\dots,\xi_n)$.

Proof. Consider the Fourier transform of identity (i). Then we have

$$\xi_i\hat{F}_j=\xi_j\hat{F}_i \quad \text{for all } i\neq j.$$

Multiply the equality below by $\xi_j \bar{\hat{F}}_i$, and adding for $i \neq j$, we have

$$\sum_{\substack{i,j=1\\ i\neq j}}^{n} \operatorname{Re} \xi_i\xi_j \hat{F}_i \bar{\hat{F}}_j = \sum_{\substack{i,j=1\\ i\neq j}}^{n} \xi_j^2 |\hat{F}_i|^2.$$

Using this identity we conclude that

$$\left|\sum_{i=1}^{n} \xi_i \hat{F}_i\right|^2 = \sum_{i=1}^{n} \xi_i^2 |\hat{F}_i|^2 + \sum_{\substack{i,j=1\\ i\neq j}}^{n} \operatorname{Re} \xi_i\xi_j \hat{F}_i \bar{\hat{F}}_j$$

$$= \sum_{i=1}^{n} \xi_i^2 |\hat{F}_i|^2 + \sum_{\substack{i,j=1\\ i\neq j}}^{n} \xi_j^2 |\hat{F}_i|^2 = \left\{\sum_{i=1}^{n} \xi_i^2\right\}\left\{\sum_{i=1}^{n} |\hat{F}_i|^2\right\}$$

which proves Lemma 3.4. □

Let us define the energy associated to system (2.1)-(2.3) as

$$E_1(t) = \frac{1}{2}\int_{\mathbb{R}^n} \left[|\mathbf{u}_t|^2 + (a^2 - b^2)\sum_{i=1}^{n} |\nabla u_i|^2 + b^2 |\operatorname{div} \mathbf{u}|^2 + \frac{\alpha}{\beta}|\theta|^2\right] dx.$$

Applying Plancherel's identity to the energy function, we have

$$E_1(t) = \frac{1}{2}\int_{\mathbb{R}^n} \left[|\hat{\mathbf{u}}_t|^2 + (a^2 - b^2)\sum_{i=1}^{n} |\xi|^2 |\hat{u}_i|^2 + b^2 \left|\sum_{i=1}^{n} \xi_i \hat{u}_i\right|^2 + \frac{\alpha}{\beta}|\hat{\theta}|^2\right] d\xi,$$

where $\xi = (\xi_1, \ldots, \xi_n)$. Let us denote by $E(\xi, t)$ the energy density given by

$$E(\xi, t) = \frac{1}{2}|\hat{\mathbf{u}}_t|^2 + (a^2 - b^2)\sum_{i=1}^{n} |\xi|^2 |\hat{u}_i|^2 + b^2 \left|\sum_{i=1}^{n} \xi_i \hat{u}_i\right|^2 + \frac{\alpha}{\beta}|\hat{\theta}|^2.$$

Let us introduce the following functions:

$$E_2(t) = \frac{1}{2}\int_{\mathbb{R}^n} |\xi|^2 E(\xi, t)\, d\xi,$$

$$F(\xi, t) = \operatorname{sign}(\alpha) \operatorname{Re}\left\{i\theta(\xi, t)\sum_{i=1}^{n} \xi_i (\bar{\hat{u}}_i)_t\right\},$$

$$G(\xi, t) = \operatorname{sign}(\alpha) |\xi|^2 \operatorname{Re}\{\hat{\mathrm{u}}_t \cdot \bar{\hat{\mathbf{u}}}\}.$$

We will prove the asymptotic behaviour of the energy by studying the asymptotic properties of the total energy density $E(\xi, t)$. By taking the Fourier transform of the system (2.1)-(2.3), we have

$$\hat{\mathbf{u}}_{tt} + (a^2 - b^2)|\xi|^2 \hat{\mathbf{u}} + b^2\left(\sum_{i=1}^{n} \xi_i \hat{u}_i\right)\xi - \alpha i \xi \hat{\theta} = 0 \quad \text{in } \mathbb{R}^n_\xi \times \mathbb{R}_t, \tag{3.5}$$

$$\hat{\theta}_t + k|\xi|^2 \hat{\theta} - \beta i \sum_{i=1}^{n} \xi_i \{\hat{u}_i\}_t = 0 \quad \text{in } \mathbb{R}^n_\xi \times \mathbb{R}_t. \tag{3.6}$$

Under these conditions we have the following lemma.

LEMMA 3.5. *With the same hypotheses of Lemma* 3.3 *we have that the derivatives of the functions E, F, and G satisfy the following inequalities*:

$$\frac{d}{dt} E(\xi, t) = -k\frac{\alpha}{\beta}|\xi|^2 |\hat{\theta}|^2, \tag{3.7}$$

$$\begin{aligned}\frac{d}{dt}F(\xi, t) &\leqq -\beta|\xi|^2|\hat{\mathbf{u}}_t|^2 + k|\xi|^2|\hat{\theta}|\{|\xi|\,|\hat{\mathbf{u}}_t|\} \\ &\quad + (a^2-b^2)|\xi|\,|\hat{\theta}|\{|\xi|^2|\hat{\mathbf{u}}|\}\end{aligned} \tag{3.8}$$

$$\frac{d}{dt}G(\xi, t) \leqq |\xi|^2|\hat{\mathbf{u}}_t|^2 - a^2|\xi|^4|\hat{\mathbf{u}}|^2 + \alpha\{|\xi|\,|\hat{\theta}|\}\{|\xi|^2|\hat{\mathbf{u}}|\}. \tag{3.9}$$

Proof. Without loss of generality we can suppose that $\alpha > 0$. By Remark 2.1 the functions E, F, G are differentiable. Multiplying (3.5) and (3.6) by $\bar{\hat{\mathbf{u}}}_t$ and $(\alpha/\beta)\hat{\theta}$, respectively, and adding the real part of the product results we obtain (3.7). Let us find the derivative of the function F:

$$\frac{d}{dt}F(\xi, t) = \operatorname{Re}\left\{i\theta_t \sum_{i=1}^{n} \xi_i(\bar{\hat{u}}_i)_t\right\} + \operatorname{Re}\left\{i\theta \sum_{i=1}^{n} \xi_i(\bar{\hat{u}}_i)_{tt}\right\}.$$

From (3.5) and (3.6) it follows that

$$\begin{aligned}\frac{d}{dt}F(\xi, t) &= -\operatorname{Re}\left\{ik|\xi|^2\hat{\theta} \sum_{i=1}^{n} \xi_i(\bar{\hat{u}}_i)_t\right\} - \beta\left|\sum_{i=1}^{n} \xi_i\{\hat{u}_i\}_t\right|^2 \\ &\quad -\operatorname{Re}\left\{i|\xi|^2(a^2-b^2)\theta(\xi, t) \sum_{i=1}^{n} \xi_i(\bar{\hat{u}}_i)\right\} - \alpha|\xi|^2|\hat{\theta}|^2.\end{aligned}$$

From Lemma 3.4 we conclude that the above identity yields (3.8). Finally, let's calculate the derivative of the function G; then

$$\begin{aligned}\frac{d}{dt}G(\xi, t) &= |\xi|^2|\hat{\mathbf{u}}_t|^2 + |\xi|^2\hat{\mathbf{u}}_{tt}\cdot\bar{\hat{\mathbf{u}}} \\ &= |\xi|^2|\hat{\mathbf{u}}_t|^2 - |\xi|^4(a^2-b^2)|\hat{\mathbf{u}}|^2 - b^2|\xi|^2\left|\left(\sum_{i=1}^{n} \xi_i\hat{u}_i\right)\right|^2 \\ &\quad + \alpha i|\xi|^2 \sum_{i=1}^{n} \xi_i\bar{\hat{u}}_i\hat{\theta}.\end{aligned}$$

From Lemma 3.4 we conclude (3.9). The proof is now complete. □

Let us briefly mention a technical difficulty arising in the proof of the main result of this work. It is necessary to consider separately the asymptotic behaviour of the energy density in the ball $B(0,1) = \{\xi \in \mathbb{R}^n;\ |\xi| < 1\}$ and in its complementary set. We proceed first to estimate the decay of the energy density function in $\mathbb{R}^n \backslash B(0,1)$, and in this case we will prove that it approaches zero exponentially when $t \to \infty$. Then we prove that the density energy function decays algebraically in $B(0,1)$ when t approaches infinity.

THEOREM 3.1. *Let* $(\mathbf{u}_0, \mathbf{u}_1, \theta_0) \in D(\mathscr{A})$ *such that*

$$\mathbf{u}_0, \mathbf{u}_1 \in [L^1(\mathbb{R}^n)]^n, \qquad \theta_0 \in L^1(\mathbb{R}^n) \quad \textit{and}$$

$$D_i u_o^j(x) = D_j u_o^i(x), \qquad D_i u_1^j(x) = D_j u_1^i(x) \quad \textit{for all } i \neq j,$$

where $\mathbf{u}_0 = (u_0^1, \ldots, u_0^n)$, $\mathbf{u}_1 = (u_1^1, \ldots, u_1^n)$, *and* $\alpha\beta > 0$. *Then there exist a positive constant* C *such that the solution of system* (2.1)–(2.3) *satisfies*

$$E_1(t) + E_2(t) \leqq C\{E_1(0) + E_2(0)\}t^{-n/2} \quad \textit{when } t \to +\infty. \tag{i}$$

Moreover, if there exist functions $f_0^k, f_1^k, g_0 \in L^2(\mathbb{R}^n) \cap L^1(\mathbb{R}^n)$ *such that*

$$\partial^{\alpha_k} f_0^k = u_0^k, \qquad \partial^{\beta_k} f_1^k = u_1^k \quad \textit{and} \quad \partial^{\alpha} g_0 = \theta_0 \qquad k = 1, \ldots, n. \tag{3.10}$$

Then we have

(ii) $$E_1(t)+E_2(t) \leqq C\{E_1(0)+E_2(O)\}t^{-m-n/2} \quad \textit{when } t \to +\infty,$$

where $m=\min\{|\alpha_k|+1, |\beta_k|, |\alpha|;\ k=1,\dots,n\}$.

Proof. Without loss of generality we can suppose that $\alpha, \beta>0$. Multiplying by $|\xi|^2$ relation (3.7) we obtain

$$\frac{d}{dt}|\xi|^2E(\xi,t)=-k\frac{\alpha}{\beta}|\xi|^4|\hat{\theta}|^2. \tag{3.11}$$

On the other hand from relation (3.8) it follows that

$$\begin{aligned}\frac{d}{dt}F(\xi,t) \leqq -\frac{1}{2}\beta|\xi|^2|\hat{\mathbf{u}}_t|^2+\frac{k^2}{2\beta}|\xi|^4|\hat{\theta}|^2+4\frac{a^2-b^2}{\beta a^2}|\xi|^2|\hat{\theta}|^2 \\ +\beta\frac{a^2}{16}|\xi|^4|\hat{\mathbf{u}}|^2,\end{aligned} \tag{3.12}$$

and from (3.9) we obtain

$$\frac{d}{dt}G(\xi,t) \leqq |\xi|^2|\hat{\mathbf{u}}_t|^2-\frac{1}{2}a^2|\xi|^4|\hat{\mathbf{u}}|^2+\frac{1}{2}\left(\frac{\alpha^2}{a^2}\right)|\xi|^2|\hat{\theta}|^2. \tag{3.13}$$

From (3.12) and (3.13) we conclude that

$$\begin{aligned}\frac{d}{dt}\left\{F(\xi,t)+\frac{\beta}{4}G(\xi,t)\right\} \leqq -\frac{\beta}{4}|\xi|^2|\hat{\mathbf{u}}_t|^2-\frac{\beta a^2}{16}|\xi|^4|\hat{\mathbf{u}}|^2 \\ +\frac{k^2}{2\beta}|\xi|^4|\hat{\theta}|^2+\left\{4\frac{a^2-b^2}{\beta a^2}+\frac{\beta\alpha^2}{8a^2}\right\}|\xi|^2|\hat{\theta}|^2.\end{aligned} \tag{3.14}$$

For convenience we introduce the function $H(\xi,t)$ defined as

$$H(\xi,t)=F(\xi,t)+\frac{\beta}{4}G(\xi,t)+N(1+|\xi|^2)E(\xi,t),$$

where N is a positive real number satisfying

$$N>\max\left\{\frac{k^2}{2\beta}; 4\frac{a^2-b^2}{\beta a^2}+\frac{\beta\alpha^2}{8a^2}\right\}+c_0,$$

where $c_0>0$ is chosen such that

$$H(\xi,t) \geqq (1+|\xi|^2)E(\xi,t). \tag{3.15}$$

From (3.7), (3.11), (3.14), and the definition of H and N we conclude that

$$\frac{d}{dt}H(\xi,t) \leqq -\frac{\beta}{4}|\xi|^2|\hat{\mathbf{u}}_t|^2-\frac{\beta a^2}{16}|\xi|^4-c_0(1+|\xi|^2)E(\xi,t).$$

From there it follows that there exists a positive constant c_1 satisfying

$$\frac{d}{dt}H(\xi,t) \leqq -c_1|\xi|^2E(\xi,t) \quad \forall \xi \in \mathbb{R}^n. \tag{3.16}$$

In order to prove part (i) we will consider two cases. First we will suppose that $|\xi| \geqq 1$. In this case the following inequalities are valid:

$$G(\xi, t) = \operatorname{Re}\{|\xi|^2 \hat{\mathbf{u}}_t \bar{\hat{\mathbf{u}}}\} \leqq \tfrac{1}{2}\{|\xi|^2|\hat{\mathbf{u}}_t|^2 + |\xi|^4|\hat{\mathbf{u}}|^2\} \quad \forall|\xi| \geqq 1,$$

$$F(\xi, t) = \operatorname{Re}\{i\xi\hat{\theta}\bar{\hat{\mathbf{u}}}_t\} \leqq \tfrac{1}{2}\{|\xi|^2|\hat{\mathbf{u}}_t|^2 + |\xi|^2|\hat{\theta}|^2\} \quad \forall|\xi| \geqq 1.$$

Thus, there exists a positive constant c_2 such that

$$H(\xi, t) \geqq c_2|\xi|^2 E(\xi, t), \tag{3.17}$$

which, together with (3.16), yields

$$\frac{d}{dt} H(\xi, t) + \gamma H(\xi, t) \leqq 0,$$

where $\gamma = c_1/c_2$. The last inequality, together with relation (3.15), implies that the energy density function satisfies

$$(1+|\xi|^2)E(\xi, t) \leqq H(\xi, t) \leqq H(\xi, 0)e^{-\gamma t} \quad |\xi| \geqq 1. \tag{3.18}$$

Integrating (3.18) in $\mathbb{R}^n \backslash B(0, 1)$ implies that

$$\int_{|\xi| \geqq 1} (1+|\xi|^2)E(\xi, t)\, d\xi \leqq c_3 E_2(0)e^{-\gamma t}. \tag{3.19}$$

Finally we will show that the asymptotic form of the density energy integral for the case $|\xi| \leqq 1$ decays algebraically, and the rate of decay increases with the increasing values of m defined in Theorem 3.1. First notice that the inequality

$$H(\xi, t) \leqq c_4 E(\xi, t) \quad \forall|\xi| \leqq 1 \tag{3.20}$$

holds. From (3.16) and (3.20) we obtain for $\gamma' = c_1/c_4$ that

$$\frac{d}{dt} H(\xi, t) + \gamma'|\xi|^2 H(\xi, t) \leqq 0,$$

which, together with (3.15), implies that

$$(1+|\xi|^2)E(\xi, t) \leqq H(\xi, t) \leqq H(\xi, 0)e^{-\gamma'|\xi|^2 t}. \tag{3.21}$$

Integrating over $B(0, 1)$ the last relation and applying Lemma 3.2, we conclude the validity of the inequality

$$\int_{|\xi| \leqq 1} (1+|\xi|^2)E(\xi, t)\, d\xi \leqq 2c_5 \int_{|\xi| \leqq 1} e^{-\gamma|\xi|^2 t}\, d\xi \leqq c_6 t^{-n/2} \tag{3.22}$$

because $H(\xi, 0)$ is bounded in $B(0, 1)$ since $\mathbf{u}_0, \mathbf{u}_1 \in [L^1(\mathbb{R}^n)]^n$, and $\theta_0 \in L^1(\mathbb{R}^n)$. From (3.19) and (3.22) part (i) follows. Finally in order to prove part (ii), let us suppose that there exist functions $f_0^k, f_1^k, g_0^k \in L^2(\mathbb{R}^n) \cap L^1(\mathbb{R}^n)$ satisfying (3.10). Then it follows that there exists a positive constant C satisfying

$$H(\xi, 0) \leqq C\{|\hat{\mathbf{u}}_1(\xi)|^2 + |\xi|^2|\hat{\mathbf{u}}_0(\xi)|^2 + |\hat{\theta}(\xi)|^2\}.$$

From 3.10 and Lemma 3.1 we have that

$$H(\xi, 0) \leqq C \sum_{k=1}^{n} |\xi^{\alpha_k}| \int_{\mathbb{R}^n} |f_0^k|\, d\xi + |\xi|^2 \sum_{k=1}^{n} |\xi^{\beta_k}| \int_{\mathbb{R}^n} |f_1^k|\, d\xi$$
$$+ \sum_{k=1}^{n} |\xi^{\alpha}| \int_{\mathbb{R}^n} |\mathscr{E}^k|\, d\xi,$$

from where it follows that there exists a positive constant C_2 such that

$$H(\xi,0)\leqq C_2\left[\sum_{k=1}^{n}|\xi^{\alpha_k}|+|\xi|^2\sum_{k=1}^{n}|\xi^{\beta_k}|+\sum_{k=1}^{n}|\xi^{\alpha}|\right],$$

where

$$C_2=\max\left\{\int_{\mathbb{R}^n}|f_0^k|\,d\xi,\int_{\mathbb{R}^n}|f_1^k|\,d\xi,\int_{\mathbb{R}^n}|g^k|\,d\xi;\,k=1,\dots,n\right\}.$$

Since

$$\begin{aligned}&\int_{|\xi\gamma\leqq 1}(1+|\xi|^2)E(\xi,t)\,d\xi\\&\leqq\int_{|\xi|\leqq 1}H(\xi,0)e^{-\gamma|\xi|^2t}\\&\leqq C_2\int_{|\xi|\leqq 1}\left[\sum_{k=1}^{n}|\xi^{\alpha_k}|+|\xi|^2\sum_{k=1}^{n}|\xi^{\beta_k}|+\sum_{k=1}^{n}|\xi^{\alpha}|\right]e^{-\gamma|\xi|^2t}\,d\xi\end{aligned}$$

by iterated integration and applying Lemma 3.2 we conclude that

$$\int_{|\xi|\leqq 1}(1+|\xi|^2)E(\xi,t)\,d\xi\leqq c_6t^{-m-n/2}\tag{3.23}$$

for some positive constant c_6. The conclusion of the theorem follows immediately from (3.19), (3.23). □

For the one-dimensional case there is no restriction on the initial data, that is, the system is totally dissipative in the sense that the total energy decays, as indicated in the following corollary.

COROLLARY 3.1. *Let* $(u_0,u_1,\theta_0)\in D(\mathscr{A})$ *such that* $u_0,u_1,\theta_0\in L^1(\mathbb{R}^1)$ *and* $\alpha\beta>0$. *Then there exist a positive constant C such that the solution of system* (1.1)-(1.3) *satisfies*

$$\text{(i)}\qquad E_1(t)+E_2(t)\leqq C\{E_1(0)+E_2(0)\}\frac{1}{\sqrt{t}}\quad\textit{when } t\to+\infty.$$

Moreover, if there exist functions $f_0,f_1,g_0\in L^2(\mathbb{R}^n)\cap L^1(\mathbb{R}^n)$ *such that*

$$\frac{d^r}{dx^r}f_0=u_0,\qquad\frac{d^s}{dx^s}f_1=u_1\quad\textit{and}\quad\frac{d^l}{dx^l}g_0=\theta_0,$$

then we have

$$\text{(ii)}\qquad E_1(t)+E_2(t)\leqq C\{E_1(0)+E_2(0)\}\left(\frac{1}{t}\right)^{m+1/2}\quad\textit{when } t\to+\infty,$$

where $m=\min\{r,s,l\}$.

Proof. For this case the energy of the system is given by

$$E_1(t)=\frac{1}{2}\int_{\mathbb{R}}|u_t(x,t)|^2+a^2|u_x(x,t)|^2+\frac{\alpha}{\beta}|\theta(x,t)|^2\,dx,$$

and the energy density is given by

$$E(\eta,t)=\frac{1}{2}\left\{|\hat{u}_t(\eta,t)|^2+a^2\eta^2|\hat{u}(\eta,t)|^2+\frac{\alpha}{\beta}|\hat{\theta}(\eta,t)|^2\right\}.$$

The auxiliary functions are

$$F(\eta, t) = \text{sign}\,(\alpha)\ \text{Re}\,\{i\eta\hat{\theta}(\eta, t)\bar{\hat{u}}_t(\eta, t)\},$$
$$G(\eta, t) = \text{sign}\,(\alpha)\ \text{Re}\,\{\eta^2\hat{u}_t(\eta, t)\bar{\hat{u}}(\eta, t)\}.$$

Finding the derivative of F and G gives

$$\frac{d}{dt}E(\eta, t) = -\frac{\alpha}{\beta}k^2\eta^2|\hat{\theta}|^2,$$

$$\frac{d}{dt}F(\eta, t) \leqq -\beta\eta^2|\hat{u}_t|^2 + k\eta^2|\hat{\theta}|\{\eta|\hat{u}_t|\} + a^2\eta|\hat{\theta}|\{\eta^2|\hat{u}|\},$$

$$\frac{d}{dt}G(\eta, t) \leqq |\hat{u}_t|^2 - a^2\eta^2|\hat{u}|^2 + \alpha\eta|\hat{\theta}|\,|\hat{u}|\}.$$

From here on we can repeat the proof as in Theorem 2.1. □

4. Decomposition of the displacement field. In this section we will study the decomposition of the displacement vector field in $\mathbb{R}^n$. We will prove that it can be decomposed into two parts, one of them the solenoidal part that conserves its energy and the other, a gradient (the irrotational part for $n = 3$) that decays as indicated in Theorem 3.1. In order to assist the reader we will show the conditions for which such decomposition holds for any vector field of $\mathbb{R}^n$.

Remark 4.1. Let us denote by $U(x)$ the function defined by

$$U(\mathbf{x}) = \frac{1}{2\pi}\ln|\mathbf{x}| \quad \text{if } n = 2,$$

$$U(\mathbf{x}) = -\frac{1}{(n-2)\sigma_n|\mathbf{x}|^{n-2}} \quad \text{if } n > 2,$$

where σ_n is the area of the unitary ball of $\mathbb{R}^n$. The solution of the problem

$$\Delta u = f \quad \text{in } \mathbb{R}^n \tag{4.1}$$

is given by

$$u(\mathbf{x}) = \int_{\mathbb{R}^n} U(\mathbf{x} - \xi)f(\xi)\,d\xi,$$

whenever f is a continuous function with compact support in $\mathbb{R}^n$. First note that for any v, and $\mathbf{C}^2$-function with compact support, the following identity is valid:

$$v(x) = \int_{\mathbb{R}^n} U(\mathbf{x} - \xi)\Delta v(\xi)\,d\xi \quad \text{in } \mathbb{R}^n.$$

Now we will prove that u satisfies (4.1). In fact, let ϕ be a $C^2(\mathbb{R}^n)$ function with compact support in $\mathbb{R}^n$. Then we have that

$$\begin{aligned}\int_{\mathbb{R}^n} \Delta u(\mathbf{x})\phi(\mathbf{x})\,d\mathbf{x} &= \int_{\mathbb{R}^n} \Delta\phi(\mathbf{x}) \int_{\mathbb{R}^n} U(\mathbf{x} - \xi)f(\xi)\,d\xi\,d\mathbf{x} \\ &\int_{\mathbb{R}^n} f(\xi) \int_{\mathbb{R}^n} U(\mathbf{x} - \xi)\Delta\phi(\mathbf{x})\,d\mathbf{x}\,d\xi \\ &= \int_{\mathbb{R}^n} f(\xi)\phi(\xi)\,d\xi \quad \forall\phi.\end{aligned}$$

By DuBois Raymond's lemma our assertion follows.

In the following lemma we will establish some regularity properties of the solution of (4.1).

LEMMA 4.1. *Let f be a function satisfying*

$$f \in C(\mathbb{R}^n)\ o(f) = o(|\mathbf{x}|^{-\theta}) \quad \text{when } |\mathbf{x}| \to +\infty;\ \theta > 2 \quad \text{for } n = 2.$$

Then there exists a continuous function u with $\partial^\alpha u \in H^1(\mathbb{R}^n)$ for $|\alpha| = 1$, satisfying (4.1). *Finally if*

$$f \in L^1(\mathbb{R}^n) \cap L^p(\mathbb{R}^n) \quad \text{for } n \geqq 3,$$

where $p \in \max\{q, q'\}$, $q > n/n-2$ and $(1/q)+(1/q') = 1$. Then there exists a solution u of (4.1) *satisfying $u \in L^q(\mathbb{R}^n)$; $\partial^\alpha u \in H^1(\mathbb{R}^n)|\alpha| = 1$.*

Proof. Let us denote by f_ν the regularization of f, that is, the convolution

$$f_\nu = \rho_\nu * f,$$

where ρ_ν is the mollifier, taken such that $\rho_\nu(-x) = \rho_\nu(x)$, $\int_{\mathbb{R}^n} \rho_\nu(\xi)\, d\xi = 1$, and $\rho_\nu(x) = 0$ if $|x| \leqq 1/\nu$. It is well known that f_ν converge to f in L^r for any $r \geqq 1$. Then the sequence $(u_\nu)_{\nu \in \mathbb{N}}$ given by

$$u_\nu(\mathbf{x}) = \int_{\mathbb{R}^n} U(\mathbf{x} - \xi) f_\nu(\xi)\, d\xi$$

satisfies the equation

$$\Delta u_\nu = f_\nu. \tag{4.2}$$

First we will consider the case $n = 2$. For this case the derivative of the function u_ν satisfies

$$\left|\frac{\partial}{\partial x_i} u_\nu(\mathbf{x})\right| \leqq \frac{1}{2\pi} \int_{\mathbb{R}^n} |\mathbf{x} - \xi|^{-1} |f_\nu(\xi)|\, d\xi.$$

If we denote by χ and χ_c the characteristic function on the open ball $B(0, 1)$ and its complementary, respectively, we have that

$$\int_{\mathbb{R}^n} |\mathbf{x} - \xi|^{-1} |f_\nu(\xi)|\, d\xi = [\chi|\xi|^{-1}] * |f_\nu| + [\chi_c|\xi|^{-1}] * |f_\nu|.$$

By the hypotheses on f we conclude that $f \in L^r(\mathbb{R}^n)$ for all $r \geqq 1$. From the fact that $\theta > 2$, there exist $p > 2$ such that $\theta - (1 - (2/p)) > 2$. Since $\chi|\xi|^{-1} \in L^1(\mathbb{R}^n)$ and $\chi_c|\xi|^{-1} \in L^p(\mathbb{R}^n)$, we have that

$$[\chi|\xi|^{-1}] * |f_\nu| \in L^p(\mathbb{R}^n) \quad \text{and} \quad [\chi_c|\xi|^{-1}] * |f_\nu| \in L^p(\mathbb{R}^n).$$

Therefore,

$$\frac{\partial}{\partial x_i} u_\nu(x) \in L^p(\mathbb{R}^n)$$

and

$$\int_{\mathbb{R}^n} \left|\frac{\partial}{\partial x_i} u_\nu(x)\right|^p d\mathbf{x} \leqq \frac{1}{2\pi} \int_{\mathbb{R}^n} \chi|\xi|^{-1}\, d\mathbf{x} \int_{\mathbb{R}^n} |f_\nu|^p\, d\mathbf{x} + \frac{1}{2\pi} \int_{\mathbb{R}^n} \chi_c|\xi|^p\, d\mathbf{x} \int_{\mathbb{R}^n} |f_\nu|\, d\mathbf{x}.$$

With the same reasoning we can conclude that

$$\int_{\mathbb{R}^n} \left|\frac{\partial}{\partial x_i} u_\nu(\mathbf{x}) - \frac{\partial}{\partial x_i} u_\mu(\mathbf{x})\right|^p d\mathbf{x}$$

$$\leqq \frac{1}{2\pi} \int_{\mathbb{R}^n} \chi|\xi|^{-1}\, d\mathbf{x} \int_{\mathbb{R}^n} |f_\nu - f_\mu|^p\, d\mathbf{x} + \frac{1}{2\pi} \int_{\mathbb{R}^n} \chi_c|\xi|^{-p}\, d\mathbf{x} \int_{\mathbb{R}^n} |f_\nu - f_\mu|\, d\mathbf{x},$$

hence $(\partial u_\nu/\partial x_i)_{\nu\in\mathbb{N}}$ is a Cauchy sequence $L^p(\mathbb{R}^n)$ then bounded. Since

$$u_\nu(\xi)-u_\nu(\mathbf{x})=\int_0^1\frac{d}{dt}u_\nu(t\xi-(1-t)\mathbf{x})\,dt$$

it follows that

$$u_\nu(\xi)-u_\nu(\mathbf{x})=\int_0^1\nabla u_\nu(t\xi+(1-t)\mathbf{x})\cdot[\xi-\mathbf{x}]\,dt.$$

Integrating over the ball B of center $\mathbf{x}$ and radius $\frac{1}{2}|\mathbf{x}-\mathbf{y}|$,

$$\begin{aligned}\left|\int_B u_\nu(\xi)\,d\xi-\frac{\pi}{4}|\mathbf{x}-\mathbf{y}|^2u_\nu(\mathbf{x})\right| &\leqq\int_B\int_0^1|\nabla u_\nu(t\xi+(1-t)\mathbf{x})\cdot[\xi-\mathbf{x}]|\,dt\,d\xi\\ &\leqq|\mathbf{x}-\mathbf{y}|\int_B\int_0^1|\nabla u_\nu(t(\xi-\mathbf{x})+\mathbf{x})|\,dt\,d\xi.\end{aligned}$$

Making a change of variable we obtain that

$$\left|{}_B\, u_\nu(\xi)\,d\xi-\frac{\pi}{4}|\mathbf{x}-\mathbf{y}|^2u_\nu(\mathbf{x})\right|\leqq|\mathbf{x}-\mathbf{y}|\int_0^1t^{-2}\int_{tD}|\nabla u_\nu(\xi)|\,d\xi\,dt, \tag{4.3}$$

where D is disk of center at the origin of coordinates and radius equal to $t|\mathbf{x}-\mathbf{y}|$. Since $t\leqq 1$, we then have that the $tD\subseteq D$, from which it follows that

$$\int_{tD}|\nabla u_\nu(\xi)|\,d\xi\leqq\pi^{1/p'}|\mathbf{x}-\mathbf{y}|^{2/p'}t^{2/p'}\left(\int_D|\nabla u_\nu(\xi)|^p\,d\xi\right)^{1/p}.$$

The last inequality and (4.3) yield

$$\begin{aligned}&\left|\int_B u_\nu(\xi)\,d\xi-\frac{\pi}{4}|\mathbf{x}-\mathbf{y}|^2u_\nu(\mathbf{x})\right|\\ &\quad\leqq\pi^{1/p'}|\mathbf{x}-\mathbf{y}|^{1+2/p'}\int_0^1t^{-2+2/p'}\left(\int_D|\nabla u_\nu(\xi)|^p\,d\xi\right)^{1/p}dt,\end{aligned}$$

from where it follows that

$$\left|\int_B u_\nu(\xi)\,d\xi-\frac{\pi}{4}|\mathbf{x}-\mathbf{y}|^2u_\nu(\mathbf{x})\right|\leqq\pi^{1/p'}\frac{p}{p-2}|\mathbf{x}-\mathbf{y}|^{1+2/p'}\left(\int_D|\nabla u_\nu(\xi)|^p\,d\xi\right)^{1/p}.$$

In the same way we obtain

$$\left|\int_B u_\nu(\xi)\,d\xi-\frac{\pi}{4}|\mathbf{x}-\mathbf{y}|^2u_\nu(\mathbf{y})\right|\leqq\pi^{1/p'}\frac{p}{p-2}|\mathbf{x}-\mathbf{y}|^{1+2/p'}\left(\int_D|\nabla u_\nu(\xi)|^p\,d\xi\right)^{1/p}.$$

By the triangle inequality we obtain

$$\begin{aligned}&\frac{\pi}{4}|\mathbf{x}-\mathbf{y}|^2|u_\nu(\mathbf{x})-u_\nu(\mathbf{y})|\\ &\quad\leqq\left|\int_B u_\nu(\xi)\,d\xi-\frac{\pi}{4}|\mathbf{x}-\mathbf{y}|^2u_\nu(\mathbf{x})\right|+\left|\int_B u_\nu(\xi)\,d\xi-\frac{\pi}{4}|\mathbf{x}-\mathbf{y}|^2u_\nu(\mathbf{y})\right|,\end{aligned}$$

from where it follows that

$$|u_\nu(\mathbf{x})-u_\nu(\mathbf{y})|\leqq 8\frac{p}{p-2}|\mathbf{x}-\mathbf{y}|^{1-2/p}\left(\int_D|\nabla u_\nu(\xi)|^p\,d\xi\right)^{1/p}. \tag{4.4}$$

With the same reasoning (for $\mathbf{y}=0$ and $u_\nu - u_\mu$ on the place of u_ν) we can prove that

$$|u_\nu(\mathbf{x}) - u_\mu(\mathbf{x})| \leqq |u_\nu(0) - u_\mu(0)| \\ + 8\frac{p}{p-2}|\mathbf{x}|^{1-2/p}\left(\int_D |\nabla u_\nu(\xi) - \nabla u_\mu(\xi)|^p \, d\xi\right)^{1/p}.$$

Then in order to prove that $(u_\nu)_{\nu\in\mathbb{N}}$ is a Cauchy sequence we only need to prove that the numerical sequence $(u_\nu(0))_{\nu\in\mathbb{N}}$ is convergent, but this follows immediately by the definition of u_ν and by the hypotheses on f. Then we conclude that there exists a continuous function u such that

$$u_\nu \to u \quad \text{uniformly on bounded sets of } \mathbb{R}^2,$$

$$\partial^\alpha u_\nu \to \partial^\alpha u \quad \text{strongly in } L^p(\mathbb{R}^2)\ p>2\ |\alpha|=1.$$

From (4.4) we also conclude that

$$|u(\mathbf{x})| \leqq |u(0)| + 8\frac{p}{p-2}|\mathbf{x}|^{1-2/p}\left(\int_D |\partial(\xi)|^p \, d\xi\right)^{1/p}.$$

Since $o(f) = o(|\mathbf{x}|^{-\theta})$, then $o(f_\nu) = o(|\mathbf{x}|^{-\theta})$ (consider the identity

$$f_\nu(\mathbf{x}) = \int_{|\xi|\leqq 1/\nu} \rho_\nu(\xi) f(\mathbf{x}-\xi)\, d\xi$$

and apply the hypotheses on f). We will prove that $\partial^\alpha u \in H^1(\mathbb{R}^n)$ for $|\alpha|=1$. In fact multiplying (4.2) by u_ν and integrating in $\mathbb{R}^n$ we have

$$\int_{\mathbb{R}^n} |\nabla u_\nu(\mathbf{x})|^2 \, d\mathbf{x} \leqq \int_{\mathbb{R}^n} |u_\nu|\,|f_\nu(\mathbf{x})| \, d\mathbf{x} \\ \leqq |u_\nu(0)| \int_{\mathbb{R}^n} |f_\nu(\mathbf{x})| \, d\mathbf{x} + C\int_{\mathbb{R}^n} |\mathbf{x}|^{1-2/p}|f_\nu(\mathbf{x})| \, d\mathbf{x}.$$

By the choice of p we have

$$\int_{\mathbb{R}^n} |\mathbf{x}|^{1-2/p}|f_\nu(\mathbf{x})| \, d\mathbf{x} \\ = \int_{|\mathbf{x}|\leqq N} |\mathbf{x}|^{1-2/p}|f_\nu(\mathbf{x})| \, d\mathbf{x} + \int_{|\mathbf{x}|\geqq N} |\mathbf{x}|^{1-2/p}|f_\nu(\mathbf{x})| \, d\mathbf{x} \\ \leqq C_1 + C_2 \int_{|\mathbf{x}|\geqq N} |\mathbf{x}|^{-\theta+1-2/p} \, d\mathbf{x} \leqq C_3.$$

It follows that there exists a constant C_4 such that

$$\int_{\mathbb{R}^n} |\nabla u_\nu(\mathbf{x})|^2 \, d\mathbf{x} \leqq C_4. \tag{4.5}$$

Moreover, since $f \in L^r(\mathbb{R}^n)$ for all $r \geqq 1$, we then have that

$$\int_{\mathbb{R}^n} |\Delta u_\nu(\mathbf{x})|^2 \, d\mathbf{x} \leqq \int_{\mathbb{R}^n} |f_\nu(\mathbf{x})|^2 \, d\mathbf{x},$$

from where it follows that Δu_ν is bounded. Then there exists a positive constant, say C_5, such that

$$\int_{\mathbb{R}^n} |\Delta u_\nu(\mathbf{x})|^2 \, d\mathbf{x} \leqq C_5. \tag{4.6}$$

From (4.5) and (4.6) the result follows for case $n=2$.

Let us consider the case $n>2$. We will prove that $(\partial^\alpha u_\nu)_{\nu\in\mathbb{N}}$ is bounded in $H^1(\mathbb{R}^n)$ for $|\alpha|=1$. In fact,

$$\begin{aligned} u_\nu(\mathbf{x}) &= \int_{\mathbb{R}^n} U(\xi) f_\nu(\mathbf{x}-\xi)\, d\xi \\ &= \int_{\mathbf{B}} U(\xi) f_\nu(\mathbf{x}-\xi)\, d\xi + \int_{\mathbf{B}^c} U(\xi) f_\nu(\mathbf{x}-\xi)\, d\xi, \end{aligned}$$

where by B and B^c we are denoting the unit ball with center at the origin and its complementary set, respectively. Let us denote by χ and χ_c the characteristic functions of B and B^c, respectively. Then we have

$$\begin{aligned} u_\nu(\mathbf{x}) &= \int_{\mathbb{R}^n} \chi(\xi) U(\xi) f_\nu(\mathbf{x}-\xi)\, d\xi + \int_{\mathbb{R}^n} \chi_c(\xi) U(\xi) f_\nu(\mathbf{x}-\xi)\, d\xi \\ &= [\chi U] * f_\nu + [\chi_c U] * f_\nu. \end{aligned}$$

It is easy to see that χU and $\chi_c U$ belong to $L^1(\mathbb{R}^n)$ and $L^q(\mathbb{R}^n)$ for $q>n/(n-2)$, respectively. Then by the hypotheses on f (case $n \geqq 3$) and Young's inequality we have

$$[\chi U] * f_\nu \in L^q(\mathbb{R}^n) \quad \text{and} \quad [\chi_c U] * f_\nu \in L^q(\mathbb{R}^n),$$

and

$$\begin{aligned} \left[\int_{\mathbb{R}^n} |u_\nu(\mathbf{x})|^q \, d\mathbf{x}\right]^{1/q} &\leqq C \int_{\mathbb{R}^n} |\chi_c U| \, d\mathbf{x} \left[\int_{\mathbb{R}^n} |f_\nu|^q \, d\mathbf{x}\right]^{1/q} \\ &\quad + C\left[\int_{\mathbb{R}^n} |\chi_c U|^q \, d\mathbf{x}\right]^{1/q} \int_{\mathbb{R}^n} |f_\nu| \, d\mathbf{x}. \end{aligned}$$

In the same way we have

$$\begin{aligned} &\left[\int_{\mathbb{R}^n} |u_\nu(\mathbf{x}) - u_\mu(\mathbf{x})|^q \, d\mathbf{x}\right]^{1/q} \\ &\quad \leqq C \int_{\mathbb{R}^n} |\chi_c U| \, d\mathbf{x} \left[\int_{\mathbb{R}^n} |f_\nu - f_\mu| \, d\mathbf{x}\right]^{1/q} \\ &\quad\quad + C\left[\int_{\mathbb{R}^n} |\chi_c U|^q \, d\mathbf{x}\right]^{1/q} \int_{\mathbb{R}^n} |f_\nu - f_\mu| \, d\mathbf{x}. \end{aligned}$$

Hence $(u_\nu)_{\nu\in\mathbb{N}}$ is a Cauchy sequence. Then it follows that there exists u in L^q such that

$$u_\nu \to u \quad \text{strong in } L^q(\mathbb{R}^n).$$

Since f_ν converges to f in $L^{q'}(\mathbb{R}^n)$ we conclude that the product $f_\nu u_\nu$ converges to fu in $L^1(\mathbb{R}^n)$. Then by multiplying (4.2) by u_ν and integrating in $\mathbb{R}^n$ we conclude that

$$\partial^\alpha u_\nu \to \partial^\alpha u \quad \text{strong in } L^2(\mathbb{R}^n) \qquad |\alpha|=1.$$

Finally by the hypotheses on f and (4.2) Δu_ν is bounded in $L^2(\mathbb{R}^n)$. Then the result follows.

LEMMA 4.2. *Let* $\mathbf{F}$ *be a vector field in* $[H^k(\mathbb{R}^n)]^n$ *such that the divergence of* $\mathbf{F}$ ($\operatorname{div}\mathbf{F} = g$) *satisfies conditions of Lemma* 4.1. *Then we can decompose* $\mathbf{F}$ *into two parts, both in* $[H^k(\mathbb{R}^n)]^n$, *one of them a gradient and the other as solenoidal function* (*that is, with null divergence*).

Proof. From Lemma 4.1 there exists a function p such that $\partial^\alpha p \in H^1(\mathbb{R}^n)$ for $|\alpha| = 1$, satisfying

$$\Delta p = \operatorname{div}\mathbf{F} \quad \text{in } \mathbb{R}^n.$$

Since $\Delta p = \operatorname{div}\mathbf{F} \in H^{k-1}(\mathbb{R}^n)$ we then have that $\partial^\alpha p \in H^k(\mathbb{R}^n)$ for all $|\alpha| = 1$; moreover, we have that

$$\mathbf{F} = \nabla p + (\mathbf{F} - \nabla p).$$

Then the result follows. □

As an application of Theorem 3.1 and Lemma 4.2 we conclude that the displacement field in thermoelasticity can be decomposed into two parts, one of them in the dissipative direction given by the gradient and the other, the solenoidal part that conserves its energy. We will express this result as the following theorem.

THEOREM 4.1. *Let* $\mathbf{u}_0$ *and* $\mathbf{u}_1$ *be vector fields in* $[H^2(\mathbb{R}^n)]^n$ *and* $[H^1(\mathbb{R}^n)]^n$, *respectively, such that* $\operatorname{div}\mathbf{u}_0$ *and* $\operatorname{div}\mathbf{u}_1$ *satisfy the hypotheses of Lemma* 4.2. *Then the displacement field in* $\mathbb{R}^n$, $n \geqq 2$, *can be decomposed into two parts, one of them solenoidal, which conserves its energy, and another totally dissipative, which decays as mentioned in Theorem* 3.1.

Proof. From Lemma 4.2 we can decompose the displacement field into two parts, say

$$\mathbf{u}_0 = \mathbf{u}_0^s + \mathbf{u}_0^i; \qquad \mathbf{u}_1 = \mathbf{u}_1^s + \mathbf{u}_1^i.$$

Where $\operatorname{div}(\mathbf{u}_0^s) = 0$, $\operatorname{div}(\mathbf{u}_1^s) = 0$, and $\mathbf{u}_c^i = \nabla p_0$, $\mathbf{u}_1^i = \nabla p_1$ for some functions p_0, p_1. Let us denote by $\mathbf{u}^s$ the solution of system

$$\begin{aligned} &\mathbf{u}^s_{tt} - (a^2 - b^2)\Delta\mathbf{u}^s = 0 \quad \text{in } \mathbb{R}^n_x \times \mathbb{R}_t, \\ &\mathbf{u}^s(0) = \mathbf{u}_0^s, \qquad \mathbf{u}^s_t(0) = \mathbf{u}_1^s \quad \text{in } \mathbb{R}^n{}_x. \end{aligned} \tag{4.7}$$

Taking the div operator on the above equation we conclude that

$$\operatorname{div}\mathbf{u}^s = \operatorname{div}\mathbf{u}^s_t = 0 \quad \text{in } \mathbb{R}^n_x.$$

From Theorem 2.1 there exists a solution $\mathbf{u}^i$ of system

$$\begin{aligned} \mathbf{u}^i_{tt} - (a^2 - b^2)\Delta\mathbf{u}^i - b^2\nabla(\operatorname{div}\mathbf{u}^i) + \alpha\nabla\theta = 0 \quad &\text{in } Q, \\ \theta_t - k\Delta\theta + \beta \operatorname{div}\mathbf{u}_t = 0 \quad &\text{in } Q, \\ \mathbf{u}^i(0) = \mathbf{u}^i_c, \quad \mathbf{u}^i_t(0) = \mathbf{u}^i_1, \quad \theta(0) = \theta_0 \quad &\text{in } Q. \end{aligned} \tag{4.8}$$

It is easy to see that $\mathbf{u} = \mathbf{u}^s + \mathbf{u}^i$ is the solution of system (2.1)-(2.3). By (4.7) we conclude that then solenoidal part conserves its energy, while the dissipative part, given by the solution of system 4.8, decays as indicated in Theorem 3.1; then the result follows. □

REFERENCES

[1] S. CHIRITA, *On the asymptotic partition of the energy in linear thermoelasticity.* Quart. Appl. Math., 45 (1987), pp. 327-340.

[2] D. G. COSTA AND W. A. STRAUSS, *Energy splitting*, Quart. Appl. Math., 39 (1981), pp. 351-361.

[3] C. M. DAFERMOS, *On the existence and the asymptotic stability of solution to the equations of linear thermoelasticity*, Arch. Rational Mech. Anal., 29 (1968), pp. 241-271.

[4] G. DASSIOS AND M. GRILLAKIS, *Dissipation rates and partition of energy in thermoelasticity*, Arch. Rational Mech. Anal., 87 (1984), pp. 49–91.

[5] R. J. DUFFIN, *Equipartition of energy in wave motion*, J. Math. Anal. Appl., 32 (1970), pp. 386-391.

[6] J. A. GOLDSTEIN, *An asymptotic property of solution of wave equations*, Proc. Amer. Math. Soc., 23 (1969), pp. 359-363.

[7] P. D. LAX AND R. S. PHILLIPS, *Scattering Theory*, Academic Press, New York, 1967.

[8] J. L. LIONS, *Quelques méthodes de resolution de problèmes aux limites non lineares*, Dunod Gauthier Villars, Paris, 1969.

[9] A. PAZY, *Semigroups of linear operators and applications to partial differential equations*, Springer-Verlag, New York, 1983.

[10] D. C. PEREIRA AND G. P. MENZALA, *Exponential decay of solutions to a coupled system of equations of linear thermoelasticity*, Comput. Appl. Math., 8 (1989), pp. 193–204.

[11] E. ZUAZUA, *Stability and decay for a class of non linear hyperbolic problems*, Asymp. Anal., 1 (1988), pp. 161–185.

SIAM J. MATH. ANAL.
Vol. 24, No. 2, pp. 407–435, March 1993

009

PERSISTENCE UNDER RELAXED POINT-DISSIPATIVITY (WITH APPLICATION TO AN ENDEMIC MODEL)*

HORST R. THIEME[†]

This paper is dedicated to Paul Waltman on the occasion of his 60th *birthday.*

Abstract. An approach to persistence theory is presented which focuses on the concept of *uniform weak persistence.* By using the most elementary dynamical systems concepts only, it can be shown that uniform weak persistence implies uniform strong persistence. This even holds under relaxed point dissipativity. Uniform weak persistence can be proved by the *method of fluctuation* or by analyzing the boundary flow for *acyclicity* with point dissipativity being only required in a neighborhood of the boundary. The approach is illustrated for a model describing the spread of a fatal infectious disease in a population that would grow exponentially without the disease. Sharp conditions are derived for both host and disease persistence and for host limitation by the disease.

Key words. persistence, permanence, dynamical systems, point dissipativity, compactness, acyclicity, epidemic (endemic) models, infectious diseases, endemicity, host limitation

AMS(MOS) subject classifications. 34C35, 92A15

Introduction. Persistence (or permanence) theory has developed into a mathematically fascinating area with important applications in mathematical ecology and epidemiology. It addresses the long-term survival of certain (if not all) components in ecological (or other) systems using and further developing the concepts and tools of dynamic systems theory. Fortunately enough, two papers by Hutson and Schmitt (preprint) and by Waltman (1991) will soon be available to give a survey of the theory as it stands now. See also Hofbauer and Sigmund (1988).

This paper addresses the following three points.

- Persistence theory in its present state is not only deep and powerful, but also conceptually difficult to grasp. Persistence is too relevant a topic to be left out from, let's say, an advanced course on mathematical modeling in ecology or epidemiology. In such a course we would expect that students are sufficiently familiar with differential equations, but are not necessarily experts in dynamical systems theory. So we present an approach which, at least for an important part, needs just the most elementary dynamical systems concepts.
- A standard assumption in today's persistence theory is point dissipativity, i.e., the existence of a bounded globally attracting set. This excludes the consideration of populations which, like the human, have grown and continue to grow without limits (at least in the time scale of interest). On the one hand, we try to establish persistence results for systems where some components are allowed to show unlimited growth. On the other hand, we suggest using persistence theory techniques to establish the boundedness of certain components of the system rather than assuming that all trajectories will be bounded.
- Persistence theory has so far focused rather on ecological than epidemiological models. In the epidemiology of infectious diseases persistence has two faces: persistence (or endemicity) of the disease and survival of the host population. We address both questions for host populations whose population size would increase in

*Received by the editors June 12, 1991; accepted for publication (in revised form) April 21, 1992.

[†]Department of Mathematics, Arizona State University, Tempe, Arizona 85287-1804.

absence of the disease. Further we derive conditions for the disease to limit the population growth.

Our approach is based on the concept of *uniform weak persistence*, which, to our knowledge, has been introduced by Freedman and Moson (1990) to supplement the concepts of *weak persistence*, *strong persistence*, and *uniform* (*strong*) *persistence* (with the last being close to *permanence*). We like to explain these concepts in terms of repelling sets.

We consider a metric space X with metric d. Let X be the union of two disjoint subsets X_1, X_2, and Φ a *continuous semiflow* on X_1, i.e., a continuous mapping $\Phi : [0,\infty) \times X_1 \to X_1$ with the following properties:

$$\Phi_t \circ \Phi_s = \Phi_{t+s}, \quad t, s \geq 0; \qquad \Phi_0(x) = x, \quad x \in X_1.$$

Here Φ_t denotes the mapping from X_1 to X_1 given by $\Phi_t(x) = \Phi(t,x)$. The *distance* $d(x,Y)$ of a point $x \in X$ from a subset Y of X is defined by

$$d(x,Y) = \inf_{y \in Y} d(x,y).$$

We use the same symbol d as for the metric because $d(x,y) = d(x,\{y\})$ for $x, y \in X$.

Let Y_2 be a subset of X_2.

Y_2 is called a *weak repeller* for X_1 if

$$\limsup_{t\to\infty} d(\Phi_t(x_1), Y_2) > 0 \quad \forall x_1 \in X_1.$$

Y_2 is called a *strong repeller* for X_1 if

$$\liminf_{t\to\infty} d(\Phi_t(x_1), Y_2) > 0 \quad \forall x_1 \in X_1.$$

Y_2 is called a *uniform weak repeller* for X_1 if there is some $\epsilon > 0$ such that

$$\limsup_{t\to\infty} d(\Phi_t(x_1), Y_2) > \epsilon \quad \forall x_1 \in X_1.$$

Y_2 is called a *uniform strong repeller* for X_1 if there is some $\epsilon > 0$ such that

$$\liminf_{t\to\infty} d(\Phi_t(x_1), Y_2) > \epsilon \quad \forall x_1 \in X_1.$$

Typically X_1 is open in X, and X_2 can be viewed as the "boundary" of X. The dynamical system Φ is called (*uniformly*) *weakly* or (*uniformly*) *strongly persistent* if X_2 is a (uniform) weak or (uniform) strong repeller for X_1.

Though weak persistence is of some interest also, uniform strong persistence is the desired property. Weak persistence does not exclude that certain components of the system get close to the boundary of extinction every now and then and are eventually wiped out by random effects. Most papers prove strong persistence first and show uniform strong persistence in a second step. Our strategy consists in proving uniform weak persistence rather than strong persistence as an intermediate result.

In our mathematical statements we speak about a certain set as a repeller for another set rather than say that the dynamical system is persistent. This gives us a greater terminological flexibility. For in endemic models, persistence of the host population and persistence (endemicity) of the disease are of separate interest and the study of each question requires us to choose the above sets X_1 and X_2 in a specific way. The repeller terminology has been used before, e.g., by Fonda (1988).

In §1 we give an elementary proof that uniform weak repellers are uniform strong repellers. This proof even works under relaxed point dissipativity. A similar result has been shown by Freedman and Moson (1990) under somewhat stronger assumptions. Point dissipativity can be removed completely at the cost of many technicalities: see §6.

In many epidemic models the *method of fluctuation* (see §2) presents an efficient and elementary way of proving uniform weak persistence. This method has been developed by Hirsch, Hanisch, and Gabriel (1985) to study global stability in certain monotone dynamical systems modeling the spread of parasitic diseases, but I feel that its main destination is persistence theory. The fluctuation method also has the advantage of providing some quantitative information.

In order to find sharp conditions for uniform weak persistence it may become necessary to study the "boundary flow" in the repeller-to-be, X_2 (§4). Then a deep plunge into dynamical systems theory becomes unavoidable, and we have to use an *acyclicity* consideration that is very similar to that needed for proving strong persistence (see Hutson and Schmitt and Waltman (1991) and the references therein, in particular Hale and Waltman (1989)). In proving uniform weak persistence, however, we are in the advantageous position that we can assume that the semiflow stays in a small neighborhood of the repeller-to-be, X_2. As already noticed by Freedman and Moson (1990), this means, e.g., that, if X_2 happens to be bounded, point dissipativity is for free even if X is unbounded. We have not explored whether there is a difference in proving uniform weak persistence as compared to strong persistence using *average Lyapunov functions* (Fonda (1988), Hofbauer and Sigmund (1988), Hutson and Schmitt).

We illustrate our approach by applying it to a model describing the spread of an infectious disease in a population that would grow without the disease. A special case of this model has been suggested by Anderson and May (1979) to study the question under which conditions infectious diseases can regulate population growth. With a different infection law, variants of this model have been studied by Busenberg and van den Driessche (1990) and Busenberg and Hadeler (1990). We introduce a contact law that includes the ones used in the papers just mentioned. We establish conditions for the disease and/or the host population to persist, as well as for the disease to limit the growth of the population. In §3 we illustrate how far one can get in this discussion with combining the method of fluctuations and the elementary persistence theory in §1. In §5 we study a situation where one has to analyze the boundary flow. Section 7 discusses a case where one has to work with almost no point dissipativity at all. Other aspects of this model like the existence and stability of disease-free and endemic equilibrium states are discussed elsewhere (see Thieme).

Although all the examples presented here are finite-dimensional, we have stated and proved our theorems to apply in infinite dimensions as well. This requires some compactness assumptions that are automatically satisfied in finite dimensions, but adds little complication to the arguments. If point dissipativity is added, our assumptions are essentially the same as in Hale and Waltman (1989). Applicable hypotheses that guarantee these compactness assumptions without point dissipativity will be derived in a subsequent publication.

1. Uniform weak repellers are uniform strong repellers. In this section we prove a persistence result, which, in the language of differential equations, only uses that solutions are continuous in time and depend continuously on their initial data. For the formulation and proof of our result we use the dynamical systems framework, however, because this is the most efficient and beautiful way of doing it. But everything can easily be reformulated in terms of differential equations.

We consider a metric space X with metric d which is the disjoint union of an open subset X_1 and a closed subset X_2. The results of this section state that, under natural conditions, X_2 is a uniform strong repeller for X_1 whenever it is a uniform weak repeller (for definitions, see the Introduction). This has already been observed by Freedman and Moson (1990) in a slightly more restrictive setting than ours. Note that the semiflow does not need to be defined on the whole state space, but on X_1 only. This will allow us to use persistence theory methods to prove boundedness results by letting X_2 contain infinity in an appropriate way. Accordingly, our assumptions will only concern the behavior of the semiflow on X_1. If Φ is defined on the whole state space X, we require that X_1 rather than X_2 is forward invariant. It is sometimes easier, however, to check certain properties of the flow (like forward invariance) on X_2 rather than on X_1.

Remark 1.1. Let the semiflow Φ be defined and continuous on all of $[0,\infty) \times X$. Then the following hold:

(a) If X_2 is a weak repeller for X_1 and X_2 is forward invariant, then X_1 is forward invariant.

(b) If X_2 is forward invariant, but not closed (i.e., X_1 is not open), then X_2 is not a weak repeller for X_1.

Remark 1.1 implies that it is natural to assume that X_1 is forward invariant and open in X. As with all persistence results, this approach requires some compactness to be associated with the system; compare Freedman and Moson (1990). But compactness is only needed in a neighborhood of the repeller-to-be. First we use the following assumption, which may appear somewhat technical:

($\mathbf{C}_{1.1}$) There is some $\delta > 0$ with the following property: If $0 < \epsilon \leq \delta$ and $x_j \in X_1, t_j > 0$, are sequences such that $t_j \to \infty, j \to \infty$, and $\mathrm{d}(\Phi_{t_j}(x_j), X_2) = \epsilon$ for all j, then the sequence $\Phi_{t_j}(x_j)$ has a convergent subsequence.

PROPOSITION 1.2. *Let X be a metric space with metric d. Let X be the disjoint union of two sets X_1 and X_2, with X_1 open. Let Φ be a continuous semiflow on X_1 such that the compactness assumption* ($\mathbf{C}_{1.1}$) *is satisfied.*

Then X_2 is a uniform strong repeller for X_1 whenever it is a uniform weak repeller for X_1.

Proof. Assume that X_2 is a uniform weak repeller. Then there exists $\epsilon > 0$ such that

$$\limsup_{t \to \infty} d(\Phi_t(x), X_2) > \epsilon \quad \forall x \in X_1. \tag{1.1}$$

We can assume that $\epsilon < \delta$ in assumption ($\mathbf{C}_{1.1}$).

Now suppose that X_2 is not a uniform strong repeller for X_1. Then there exist sequences $x_j \in X_1$ and $0 < \epsilon_j \to 0$ such that

$$\liminf_{t \to \infty} d(\Phi_t(x_j), X_2) < \epsilon_j \quad \forall j. \tag{1.2}$$

We now find sequences $r_j < s_j < t_j$ with $r_j \to \infty, j \to \infty$ such that

$$d(\Phi_{s_j}(x_j), X_2) \to 0, \qquad j \to \infty, \tag{1.3}$$

$$d(\Phi_{r_j}(x_j), X_2) = \epsilon = d(\Phi_{t_j}(x_j), X_2), \tag{1.4}$$

$$d(\Phi_s(x_j), X_2) \leq \epsilon, \qquad r_j \leq s \leq t_j. \tag{1.5}$$

Moreover, we can choose any of the numbers $\tau_j \in \{r_j, s_j, t_j\}$ in the form $\tau_j = \rho_j + \sigma_j$ such that $\sigma_j \to \infty, j \to \infty$, and both

$$(1.6) \qquad d(x_j, X_2) \leq \epsilon, \qquad d(\Phi_{\tau_j}(x_j), X_2) \leq \epsilon$$

for all j. Hence, by our compactness condition, we can assume that, after taking a subsequence, the sequence $\Phi_{r_j}(x_j)$ converges.

Step 1. We claim that $t_j - r_j$ is unbounded.

We suppose that $t_j - r_j$ is bounded. Then $s_j - r_j \to s$ after choosing a subsequence. Let $\Phi_{r_j}(x_j) \to x$. As $d(x, X_2) = \epsilon$ by (1.4), we have $x \in X_1$. On the other hand, we have $\Phi_{s_j}(x_j) = \Phi_{s_j - r_j}(\Phi_{r_j}(x_j)) \to \Phi_s(x)$. As $x \in X_1$ and X_1 is forward invariant, $\Phi_s(x) \in X_1$. As $d(\Phi_s(x), X_2) = 0$ by (1.3) and as X_2 is closed, $\Phi_s(x) \in X_2$, a contradiction. This finishes Step 1.

Step 2. The contradiction.

Let x be the limit of $\Phi_{r_j}(x_j), j \to \infty$. As $d(x, X_2) = \epsilon$ by (1.4) and as $X = X_1 \cup X_2$, we have $x \in X_1$. By (1.5),

$$d(\Phi_r \circ \Phi_{r_j}(x_j), X_2) \leq \epsilon, \qquad 0 \leq r \leq t_j - r_j.$$

By Step 1, $t_j - r_j$ is unbounded. After passing to subsequences we can assume that $t_j - r_j$ is increasing and converges to infinity. Then

$$d(\Phi_r \circ \Phi_{r_k}(x_k), X_2) \leq \epsilon, \qquad 0 \leq r \leq t_j - r_j, \quad k \geq j.$$

By the continuity of Φ we can take the limit for $k \to \infty$:

$$d(\Phi_r(x), X_2) \leq \epsilon, \qquad 0 \leq r \leq t_j - r_j.$$

As $t_j - r_j \to \infty, j \to \infty$, this estimate holds for all $r \geq 0$, contradicting $x \in X_1$ and (1.1).

If X is a subset of $\mathbf{R}^n$ with the induced metric, or, more generally, a locally compact metric space, the somewhat technical condition $(\mathbf{C}_{1.1})$ can be replaced by requiring X_2 to be compact. This generalizes a result by Freedman and Moson (1990) who assume that the semiflow is defined on the whole state space and leaves X_2 forward invariant (see our Remark 1.1a). Our version allows to prove boundedness of trajectories by making X_2 contain infinity in an appropriate way; see §3.3 for an example.

THEOREM 1.3. *Let X be a locally compact metric space with metric d. Let X be the disjoint union of two sets X_1 and X_2 such that X_2 is compact. Let Φ be a continuous semiflow on X_1.*

Then X_2 is a uniform strong repeller for X_1, whenever it is a uniform weak repeller for X_1.

Proof. If X_2 is compact and X is locally compact, X_2 has a compact neighborhood. Hence $(\mathbf{C}_{1.1})$ is satisfied if δ is chosen sufficiently small.

If X is not locally compact, Proposition 1.2 can be improved by adapting the idea of a compact attracting set. As we do not make explicit use of this concept, we refer the reader to Hale and Waltman (1989), Hutson and Schmitt, and Waltman (1991). We assume that a neighborhood of X_2 is attracted to a certain set B. Point dissipativity would require that B is bounded. For our purposes it is sufficient that the intersection of B with some neighborhood of X_2 is bounded. Then we assume that the semiflow is compactifying on bounded sets.

($\mathbf{C}_{1.2}$) There are $\delta > 0$ and a subset B of X_1 with the following properties:

- If $x \in X_1$ and $d(x, X_2) < \delta$, then $\Phi_t(x) \in B$ for all sufficiently large t;
- If $0 < \epsilon < \delta$, the intersection of B with the ϵ-shell $\{x \in X; d(x, X_2) = \epsilon\}$ of X_2 is bounded.
- If $t_j \to \infty$ and x_j is a sequence in X_1 such that the sequence $\Phi_{t_j}(x_j)$ is bounded, then $\Phi_{t_j}(x_j)$ has a convergent subsequence in X.

Remark. The third condition in ($\mathbf{C}_{1.2}$) is satisfied if there exists some $t_0 > 0$ with the following two properties:

(i) Φ_{t_0} is compact, i.e., it maps bounded sets onto relatively compact sets;

(ii) For any bounded set B_1 there exists another bounded set B_2 such that $\Phi_{t_0}^{-1}(B_1) \subset B_2$.

THEOREM 1.4. *Let X be a metric space with metric d. Let X be the disjoint union of two sets X_1 and X_2, with X_1 open. Let Φ be a continuous semiflow on X_1 such that the compactness assumption* ($\mathbf{C}_{1.2}$) *is satisfied.*

Then X_2 is a uniform strong repeller for X_1, whenever it is a uniform weak repeller for X_1.

Proof. Let us revisit the proof of Proposition 1.2. Consider (1.1)–(1.6). As we can always replace x_j by $\Phi_{\tau_j}(x_j)$ with appropriate, arbitrarily large τ_j, the first assumption in ($\mathbf{C}_{1.2}$) allows us to assume that all x_j and their forward orbits are in the set B. By (1.4), we have that the sequence $\Phi_{r_j}(x_j)$ is contained in the intersection of B and the ϵ shell of X_2 that is bounded by the second assumption in ($\mathbf{C}_{1.2}$). The third assumption then implies that, after taking a subsequence, the sequence $\Phi_{r_j}(x_j)$ converges. Then we can proceed exactly as in the proof of Proposition 1.2.

Apparently there are two routes to uniform strong persistence. The first, traditional one, leads over proving strong persistence, the second, just established, uses uniform weak persistence as an intermediate result. As far as we understand the literature, there are basically two methods to establish strong persistence: using *average Lyapunov functions* (Fonda (1988), Hofbauer and Sigmund (1988), Hutson and Schmitt) or using an *acyclicity* condition for the flow on X_2 (Hale and Waltman (1989), Hutson and Schmitt, Waltman, (1991)). Everybody working with Lyapunov functions knows that they are a wonderful theoretical tool, but sometimes difficult to construct in practice. In surprisingly many cases it is possible to establish uniform weak persistence by ad hoc methods. If X is a subset of $\mathbf{R}^n$, it is often possible to use the *method of fluctuation* which yields additional interesting information (about ϵ in Definition 1.1a, e.g.). This method is explained in the next section and applied in §3, where its limitations are also discovered. Sometimes the fluctuation method can be extended to infinite dimensions. In Thieme and Castillo-Chavez, uniform strong persistence for an HIV/AIDS model (Theorem 4) was established using Theorem 4.2 in Hale and Waltman (1989). The conditions of Theorem 4.2 were checked by establishing uniform weak persistence (Theorem 3b in Thieme and Castillo-Chavez). So we could have concluded uniform strong persistence from our Theorem 1.4 as well.

The other method of showing strong persistence, acyclicity, works for establishing uniform weak persistence as well. It even does so under relaxed point dissipativity because, for checking uniform weak persistence rather than strong persistence, we can completely restrict the consideration to the semiflow in a neighborhood of the repeller-to-be. This will be elaborated in §4.

2. The method of fluctuation. The *method of fluctuation* has been developed and systematically used by Hirsch, Hanisch, and Gabriel (1985) for studying the spread of parasitic diseases. It is a very convenient method for analyzing the asymptotic behavior of solutions of finite-dimensional differential equations. Hirsch, Hanisch, and Gabriel mainly use it to establish the global stability of equilibria. If used for this purpose, the advantage of the method lies in the fact that, if it works at all, it works easily. The disadvantage consists in its restriction to differential equations, the vector fields of which have to satisfy stringent monotonicity requirements. It is a powerful tool, however, to show that certain sets are uniform weak repellers.

For a real-valued function f on $[t_0, \infty)$ we define

$$f_\infty = \liminf_{t \to \infty} f(t), \qquad f^\infty = \limsup_{t \to \infty} f(t).$$

The following lemma can be found in Hirsch, Hanisch, and Gabriel (1985).

LEMMA 2.1 (fluctuation lemma). *Let $f : [t_0, \infty) \to \mathbf{R}$ be a differentiable function that has no limit for $t \to \infty$. Then there are sequences $s_n, t_n \to \infty$ with the following properties:*

$$f(s_n) \to f_\infty, \qquad f'(s_n) = 0,$$
$$f(t_n) \to f^\infty, \qquad f'(t_n) = 0$$

for $n \to \infty$. If f is twice continuously differentiable, we have in addition that

$$f''(s_n) \geq 0, \quad f''(t_n) \leq 0, \quad n \in \mathbf{N}.$$

This statement is quite intuitive. If f has no limit for $t \to \infty$, it has to oscillate between f_∞ and f^∞. So we can choose appropriate sequences of local minima $f(s_n)$ and local maxima $f(t_n)$ that have the desired properties.

PROPOSITION 2.2. *Let $f : (t_0, \infty) \to \mathbf{R}$ be bounded and continuously differentiable. Then there are sequences $s_n, t_n \to \infty$ with the following properties:*

$$f(s_n) \to f_\infty, \qquad f'(s_n) \to 0,$$

$$f(t_n) \to f^\infty, \qquad f'(t_n) \to 0$$

for $n \to \infty$.

Proof. By Lemma 2.1 we can assume that $f(t)$ has a finite limit for $t \to \infty$. If the statement of this proposition does not hold, f' does not change sign for sufficiently large t and must be bounded away from zero. But this contradicts the convergence of $f(t)$ to a finite limit.

THEOREM 2.3. *Let D be a bounded interval in $\mathbf{R}$ and $g : (t_0, \infty) \times D \to \mathbf{R}$ be bounded and uniformly continuous. Further, let $x : (t_0, \infty) \to D$ be a solution of*

$$x' = g(t, x),$$

which is defined on the whole interval (t_0, ∞). Then there exist sequences $s_n, t_n \to \infty$ such that

$$\lim_{n \to \infty} g(s_n, x_\infty) = 0 = \lim_{n \to \infty} g(t_n, x^\infty).$$

Proof. As g is bounded, x' is bounded and so x is uniformly continuous on (t_0, ∞). As g is continuous, so is $x'(t)$ on (t_0, ∞). By Proposition 2.2, we find sequences $s_n, t_n \to \infty$ with the following properties:

$$x(s_n) \to x_\infty, \qquad x'(s_n) = g(s_n, x(s_n)) \to 0,$$

$$x(t_n) \to x^\infty, \qquad x'(t_n) = g(t_n, x(t_n)) \to 0,$$

for $n \to \infty$. As g is uniformly continuous, the assertion follows.

For establishing the uniform weak repeller relation we will mainly use the following version.

COROLLARY 2.4. *Let the assumptions of Theorem* 2.3 *be satisfied. Then*

$$\text{(a)} \qquad \liminf_{t\to\infty} g(t, x_\infty) \le 0 \le \limsup_{t\to\infty} g(t, x_\infty),$$

$$\text{(b)} \qquad \liminf_{t\to\infty} g(t, x^\infty) \le 0 \le \limsup_{t\to\infty} g(t, x^\infty).$$

3. Applications of the fluctuation method to an endemic model: persistence of host and disease. Host limitation. (Act I). We consider the spread of a potentially fatal infectious disease in a host population that would increase exponentially in the absence of the disease. The question of persistence can be posed in a twofold way: persistence of the host population, i.e., the disease does not extinguish the host; and persistence (or endemicity) of the disease, i.e., the disease does not go extinct itself. Moreover, we can ask whether the infectious disease limits the growth of the host population. In this section we use the method of fluctuation to show that an appropriate set is a uniformly weak repeller, and we use Theorem 1.3 to establish that it is a uniformly strong repeller. For host and disease persistence we could alternatively use the results in Freedman and Moson (1990) in order to get from uniform weak to uniform strong persistence; for host limitation the analogous step seems to require the more general Theorem 1.3.

Our model follows Anderson and May (1979), Busenberg and van den Driessche (1990), and Busenberg and Hadeler (1990). The host population is subdivided into susceptible, S, infective, I, and recovered, R, individuals (a latent period is ignored):

$$\begin{aligned} &N = S + I + R, \\ &\frac{d}{dt}S = \beta N - \mu S - C(N)S\frac{I}{N} + \rho R, \\ &\frac{d}{dt}I = C(N)S\frac{I}{N} - (\mu + \gamma + \alpha)I, \\ &\frac{d}{dt}R = \gamma I - (\mu + \rho)R. \end{aligned} \tag{3.1}$$

β, μ are the per capita birth and mortality (without the disease) rates. ρ is the rate at which immunity is lost. γ is the rate at which individuals recover from the disease, while α is the extra per capita mortality due to the disease. All these constants are assumed to be strictly positive with the possible exception of ρ, which may also be zero in case the disease infers permanent immunity. As we are interested in the case where the population size increases exponentially in absence of the disease, we assume

$$\beta > \mu.$$

Model (3.1) is more special than the models considered by Busenberg and van den Driessche (1990) and Busenberg and Hadeler (1990) insofar as the disease only affects the mortality of infective individuals and not the mortality of recovered individuals and the fertility of infective and recovered individuals. Furthermore, vertical transmission is neglected. Diekmann and Kretzschmar (1991) combine reduced fertility of infectives with pair formation. This leads to a behavior of their model (bistability of solutions, global bifurcation of periodic solutions from a saddle connection) that is quite different from the one observed in the above-mentioned papers and also in this paper.

The key feature of this model that makes it different from the models considered by Anderson and May (1979) on one hand and Busenberg and van den Driessche (1990) and Busenberg and Hadeler (1990) on the other, is the way in which the rate $C(N)$ of effective contacts (i.e., contacts actually leading to an infection in case they occur between a susceptible and an infective individual) depends on the population size N. Anderson and May (1979) assume the classical mass action approach $C(N) = \text{const } N$, whereas Busenberg and van den Driessche (1990) and Busenberg and Hadeler (1990) assume $C(N) = \text{const}$. Anderson (1982) fits

$$C(N) = \text{const } N^{\delta}$$

to various communities and various childhood diseases and finds δ in the range between 0.03 and 0.07, i.e., the dependence of C on N is very weak for sufficiently large N. We must take into account, however, that these estimates are taken from different communities of constant size and do not necessarily reflect how the contact rate changes in a community of varying size. Using a handling time argument in modeling the contact rate for venereal diseases (in analogy to Holling's (1966, p. 11) derivation of a predator's functional response to the amount of prey) suggests the form chosen by Diekmann and Kretzschmar (1991) (see also Dietz (1982)),

$$C(N) = \frac{N}{\kappa_1 + \kappa_2 N}.$$

The handling time argument neglects, so to speak, competition in partner acquisition. Modeling the formation of short time social complexes, Heesterbeek and Metz derive an effective contact rate of the form

$$C(N) = \frac{\kappa_1 N}{1 + \kappa_2 N + \sqrt{1 + 2\kappa_2 N}}.$$

The two last forms of $C(N)$ approximate the form considered by Anderson and May (1979) at small population sizes and the form considered by Busenberg and van den Driessche (1990) and Busenberg and Hadeler (1990) at large population sizes and interpolate in between.

The importance of the contact function C came particularly into focus by modeling the spread of HIV—see, e.g., Castillo-Chavez et al. (1989 a,b), Thieme and Castillo-Chavez—and is now taken into account also in other models (Brauer (1990), (1991), Pugliese (1990)).

The solutions of (3.1) have many interesting properties, which will be discussed in detail elsewhere (Thieme): Depending on the parameters, in particular on the values of C at zero and ∞, there may be a steady state in which the disease controls the population, or there may be exponential states in which both the sizes of the total population and the infective and removed parts increase exponentially. If $C(0) > 0$, there can even exist a state in which the population size decreases exponentially, i.e., the disease has converted the originally exponentially increasing population into an exponentially decreasing population. Here we concentrate on how the values of C at zero and ∞ affect the persistence of both host and disease and the limitation of the host population by the disease.

In view of the various forms of C discussed before, we make the following assumptions.

Assumptions 3.1. (a) $C(N)$ is a continuous function of $N \geq 0$ and continuously differentiable in $N > 0$.

(b) $C(N)$ is monotone nondecreasing in N.

(c) $C(N) > 0$ if $N > 0$.

Reformulation of the model. It is convenient to reformulate the model in terms of the fractions of the susceptible, infective, and recovered parts of the population:

$$x = \frac{S}{N}, \quad y = \frac{I}{N}, \quad z = \frac{R}{N}, \tag{3.2}$$

and express (3.1) in these terms:

$$\begin{aligned} \frac{d}{dt}N &= (\beta - \mu - \alpha y)N, \\ \frac{d}{dt}x &= \beta(1-x) - (C(N) - \alpha)xy + \rho z, \\ \frac{d}{dt}y &= C(N)xy - (\gamma + \alpha + \beta)y + \alpha y^2, \\ \frac{d}{dt}z &= \gamma y - (\beta + \rho)z + \alpha yz. \end{aligned} \tag{3.3}$$

Equations (3.2) suggest

$$x + y + z = 1, \tag{3.4}$$

which allows us to get rid of either x, y, or z, but it is convenient to have all equations available. Actually, it is easy to see that the manifold $x + y + z = 1$, $x, y, z \geq 0$, is forward invariant under the solution flow of (3.3), which implies that, for any initial data satisfying (3.4), the system (3.3) has a global (in forward direction) solution satisfying (3.4).

3.1. Persistence of the host population. As a first application of the persistence theory in §1 in combination with the fluctuation method, we show that the host population persists in the original Anderson and May (1979) model, or, more generally, if $C(0) = 0$.

THEOREM 3.2. *Let $C(0) = 0, N(0) > 0$. Then the population is uniformly persistent, i.e., $\liminf_{t\to\infty} N(t) \geq \epsilon$ with $\epsilon > 0$ not depending on the initial data.*

Proof. We have to show that the set $X_2 = \{N = 0, x \geq 0, y \geq 0, z \geq 0, x+y+z = 1\}$ is a uniform strong repeller for $X_1 = \{N > 0, x \geq 0, y \geq 0, z \geq 0, x + y + z = 1\}$. As the assumptions of Theorem 1.3 are obviously satisfied, it is enough to show that X_2 is a uniform weak repeller for X_1. We consider a trajectory for which $N^\infty < \infty$ and apply Corollary 2.4(b) to the y equation in (3.3):

$$0 \leq C(N^\infty) - (\gamma + \beta)y^\infty.$$

Here we have used that $x, y \leq 1$. We solve for y^∞,

$$y^\infty \leq \frac{C(N^\infty)}{\gamma + \beta}. \tag{3.5}$$

From the N equation in (3.3) we obtain

$$\liminf_{t\to\infty} \frac{\frac{d}{dt}N}{N} \geq \beta - \mu - \alpha y^{\infty}.$$

Hence N increases exponentially unless

$$y^{\infty} \geq \frac{\beta - \mu}{\alpha}. \tag{3.6}$$

Combining (3.5) and (3.6) we obtain

$$C(N^{\infty}) \geq \frac{(\beta - \mu)(\gamma + \beta)}{\alpha}. \tag{3.7}$$

As $C(0) = 0$ and C is continuous at 0, $N^{\infty} \geq \epsilon > 0$ with ϵ not depending on the initial data.

Actually we see from (3.7) that we can relax $C(0) = 0$ and require

$$C(0) < (\beta - \mu)\frac{\gamma + \beta}{\alpha}$$

instead. But for obtaining a sharp persistence result (for the survival of the population) the combination of the persistence Theorem 1.3 and the fluctuation method is not sufficient. In the next section we will develop a more sophisticated method to prove uniform weak persistence, which, following the lines of Hale and Waltman (1989), analyzes the semiflow on the repeller to-be, X_2. This method, which uses an acyclicity requirement, will be applied to our endemic model in §5.

3.2. Persistence of the disease. As another application of Theorem 1.3 and the method of fluctuation we look for conditions under which the disease is persistent. There are different ways in which disease persistence can be interpreted in our model. We have chosen to call the disease to be persistent or *endemic* in the population, if the fraction of infective individuals, y, is bounded away from zero. As we do not exclude that the population size increases or decreases, this is not equivalent to the number of infectives being bounded away from zero. If the population dies out and the fraction of infectives remains bounded away from zero, we would still say that the disease is persistent in the population. On the other hand, we would not call the disease persistent if both the population size and the number of infectives increase, let us say, exponentially with the exponential growth rate of the number of infectives being strictly less than the exponential growth rate of the total population.

We first derive a condition for the disease to be weakly uniformly persistent.

PROPOSITION 3.3. *Let $\alpha + \beta + \gamma < C(\infty)$. Then the disease is uniformly weakly persistent in so far as*

$$y^{\infty} = \limsup_{t\to\infty} y(t) \geq \epsilon$$

with $\epsilon > 0$ being independent of the initial data, provided that $y(0) > 0$.

Actually, the proof shows that

$$y^{\infty} \geq \min\left(\frac{\beta - \mu}{\alpha}, \frac{1}{1 + \frac{\gamma+\alpha}{\beta+\rho}}\left(1 - \frac{\gamma + \alpha + \beta}{C(\infty)}\right)\right).$$

Proof. We suppose that the disease is not uniformly weakly persistent and derive a contradiction. We can assume that

$$y^\infty < \frac{\beta - \mu}{\alpha}.$$

The N equation in (3.3) implies that

$$\liminf_{t\to\infty} \frac{\frac{d}{dt}N}{N} > 0,$$

i.e., $C(N(t)) \to C(\infty), t \to \infty$. We apply Corollary 2.4(b) to the z equation in (3.3):

$$0 \leq \gamma y^\infty - (\beta + \rho) z^\infty + \alpha y^\infty.$$

Here we have used that $y \leq 1$. Solving for z^∞ we obtain

$$z^\infty \leq \frac{\gamma + \alpha}{\beta + \rho} y^\infty.$$

We substitute (3.4) into the y equation in (3.3):

$$\frac{d}{dt} y = C(N)(1 - y - z)y - (\gamma + \alpha + \beta)y + \alpha y^2.$$

Hence

$$\liminf_{t\to\infty} \frac{\frac{d}{dt}y}{y} \geq C(\infty)\left(1 - y^\infty \left[1 + \frac{\gamma + \alpha}{\beta + \rho}\right]\right) - (\gamma + \alpha + \beta).$$

Hence, if in addition,

$$y^\infty < \frac{1}{1 + \frac{\gamma+\alpha}{\beta+\rho}} \left(1 - \frac{\gamma + \alpha + \beta}{C(\infty)}\right),$$

we have

$$\liminf_{t\to\infty} \frac{\frac{d}{dt}y}{y} > 0,$$

which implies that $y(t) \to \infty, t \to \infty$, in contradiction to the fact that y is bounded by one. This shows the estimate for y^∞ mentioned after the statement of this proposition.

As Theorem 1.3 requires compactness of the repeller to-be, we are forced to restrict our consideration to the case that $C(\infty) < \infty$. Without this restriction we cannot guarantee that $C(N)$ ends in a bounded set. In §7 we prove uniform strong persistence of the disease if $C(\infty) = \infty$. This requires a modification of the persistence results in §1, which unfortunately adds a lot of technicalities.

THEOREM 3.4. *Let $\alpha + \beta + \gamma < C(\infty) < \infty$. Then*

$$\liminf_{t\to\infty} y(t) \geq \epsilon > 0, \qquad \liminf_{t\to\infty} z(t) \geq \epsilon > 0$$

with $\epsilon > 0$ being independent of the initial data, provided that $y(0) > 0$.

Proof. As we cannot exclude that $N(t)$ is unbounded, we choose $X = \{(N, x, y, z);\ 0 \leq N \leq \infty, x, y, z \geq 0, x+y+z = 1\}$. In order to make X a metric space we introduce

$$\varphi(N) = \begin{cases} \frac{N}{(1+N)}, & 0 \leq N < \infty, \\ 1, & N = \infty, \end{cases}$$

and set

$$d\Big((N_1,x_1,y_1,z_2),(N_2,x_2,y_2,z_2)\Big)=|\varphi(N_1)-\varphi(N_2)|+|x_1-x_2|+|y_1-y_2|+|z_1-z_2|.$$

If $N_0<\infty$, the semiflow $\Phi_t(N_0,x_0,y_0,z_0)$ is defined to be the solution of (3.3) at time t, for initial data N_0,x_0,y_0,z_0. If $N_0=\infty$, then $\Phi_t(\infty,x_0,y_0,z_0)=(\infty,x(t),y(t),z(t))$ with x,y,z being the solutions of the x,y,z equations in (3.3), with $C(N)$ being replaced by $C(\infty)$. It is easy to see that Φ is a continuous semiflow. Our metric makes X a compact space. In a first step we show that y is bounded away from zero, or, even more, that $X_2=\{(N,x,0,z);0\leq N\leq\infty,x,z\geq 0,x+y+z=1\}$ is a uniform strong repeller for $X_1=\{(N,x,y,z);0\leq N\leq\infty,x,z\geq 0,y>0,x+y+z=1\}$. Notice that both X_1 and X_2 are forward invariant. By Proposition 3.3, X_2 is a uniform weak repeller for X_1 and hence a uniform strong repeller by Theorem 1.3. Notice that $d((N,x,y,z),X_2)=y$. Hence y is bounded away from zero, with the bound being independent of the initial data. In order to show that the same holds for z we apply Corollary 2.4(a) to the z equation:

$$0\geq\gamma y_\infty-(\beta+\rho)z_\infty.$$

We solve for z_∞:

$$z_\infty\geq\frac{\gamma}{\beta+\rho}y_\infty.$$

This finishes the proof.

3.3. Host limitation by the disease. We conclude this section by using persistence theory for finding a condition to guarantee that the disease limits the population growth. Here we restrict ourselves to the case that $C(\infty)=\infty$. The case $C(\infty)<\infty$ requires a more sophisticated acyclicity consideration and will be dealt with in §5.

We introduce the following expression:

$$\mathcal{P}_0=\frac{\beta-\mu}{\alpha}\left(1+\frac{\gamma}{\mu+\rho}\right). \tag{3.8}$$

PROPOSITION 3.5. *Let $C(\infty)=\infty,\mathcal{P}_0<1$, and $y(0)>0$. Then*

$$\liminf_{t\to\infty}N(t)\leq\hat{N}<\infty$$

with $\hat{N}$ not depending on the initial conditions.

The following result is in itself of interest.

LEMMA 3.6. (a) *There exists unique solutions $y^\diamond>0,z^\diamond>0$ of the equations*

$$\begin{aligned}1&=y^\diamond+z^\diamond,\\0&=\gamma y^\diamond-(\beta+\rho)z^\diamond+\alpha y^\diamond z^\diamond.\end{aligned}$$

(b) *Furthermore,*

$$y^\diamond>\frac{\beta-\mu}{\alpha}\iff\mathcal{P}_0<1.$$

(c) *Finally, $z^\infty\leq z^\diamond$ for any solution of* (3.3). *In particular,*

$$1-z^\infty\geq y^\diamond>0\quad\textit{and}\quad y_\infty\geq y^\diamond-x^\infty.$$

Proof. In order to show part (a) we substitute the first equation into the second:

$$0 = \gamma(1 - z^\diamond) - (\beta + \rho)z^\diamond + \alpha(1 - z^\diamond)z^\diamond =: \varphi(z^\diamond). \tag{3.9}$$

The function φ, defined in (3.9), is strictly concave, $\varphi(0) > 0, \varphi(1) < 0$. Hence, by the intermediate value theorem, there is a solution $z^\diamond \in (0, 1)$, which is unique.

In order to show part (c) we consider the z equation in (3.3) and the relation $1 = x + y + z$ with x, y, x being nonnegative, and obtain the differential inequality

$$\frac{d}{dt}z \leq \gamma(1 - z) - (\beta + \rho)z + \alpha(1 - z)z.$$

The fluctuation Lemma 2.1 provides a sequence $t_n \to \infty$, $n \to \infty$, such that $z'(t_n) \to 0$, $z(t_n) \to z^\infty$, $n \to \infty$. Hence

$$0 \leq \gamma(1 - z^\infty) - (\beta + \rho)z^\infty + \alpha(1 - z^\infty)z^\infty = \varphi(z^\infty).$$

As the function φ is strictly negative for arguments in $(z^\diamond, 1]$, we obtain $z^\infty \leq z^\diamond$. This implies $1 - z^\infty \geq 1 - z^\diamond = y^\diamond > 0$ and $y_\infty \geq 1 - z^\infty - x^\infty \geq y^\diamond - x^\infty$.

In order to show part (b) we solve the second equation in Lemma 3.6 for $z^\diamond$,

$$z^\diamond = \frac{\gamma y^\diamond}{\beta + \rho - \alpha y^\diamond},$$

and we substitute this expression into the first equation,

$$1 = y^\diamond\left(1 + \frac{\gamma}{\beta + \rho - \alpha y^\diamond}\right). \tag{3.10}$$

We also notice the constraint

$$y^\diamond < \frac{\beta + \rho}{\alpha}; \tag{3.11}$$

otherwise $z^\diamond$ would be undefined or strictly negative. The right-hand side of (3.10) is a strictly increasing function of $y^\diamond \geq 0$ under the constraint (3.11). We observe that

$$y^\natural = \frac{\beta - \mu}{\alpha}$$

also satisfies the constraint (3.11). Hence $y^\diamond > y^\natural$ if and only if

$$1 > y^\natural\left(1 + \frac{\gamma}{\beta + \rho - \alpha y^\natural}\right) = \mathcal{P}_0,$$

by the definition of $\mathcal{P}_0$ in (3.8). This concludes the proof of Lemma 3.6.

Proof of Proposition 3.5. Assume that $C(N_\infty) > c > C(0) \geq 0$. We will show that c cannot be chosen arbitrarily large without obtaining a contradiction. The choice of c will not depend on the initial data. Consider the y equation in (3.3). Recall (3.4), $x + y + z = 1$. By Lemma 3.6 (c) we have $1 - z^\infty > 0$, and we can assume

$$x(t) \geq \epsilon - y(t)$$

for large enough t, with $\epsilon > 0$ being independent of the initial data. Hence, for sufficiently large t,

$$\frac{d}{dt}y \geq \Big(C(N)(\epsilon - y) - (\gamma + \alpha + \beta)\Big)y.$$

Choose $c > 0$ such that

$$c\epsilon > \gamma + \alpha + \beta.$$

c can be chosen independently of the initial data. As $C(N(t)) \geq c$ for all large t, we have that $y(t) \geq \delta$ for all sufficiently large t, with some $\delta > 0$, which is independent of the initial data. Now consider the x equation in (3.3):

$$\frac{d}{dt}x \leq \beta + \rho - \delta(C(N) - \alpha)x$$

for sufficiently large t. Choose $c > \alpha$ and recall $C(N_\infty) > c$. Then

$$x^\infty \leq \frac{\beta + \rho}{\delta(c - \alpha)}.$$

Choosing c large enough, we can achieve that

$$x^\infty < \eta$$

for any given arbitrarily small $\eta > 0$. By Lemma 3.6(c) we have

$$y_\infty \geq y^\diamond - x^\infty > y^\diamond - \eta.$$

As $\mathcal{P}_0 < 1$, Lemma 3.6(b) implies that

$$y^\diamond > y^\sharp := \frac{\beta - \mu}{\alpha}.$$

By choosing $\eta > 0$ small enough, i.e., by choosing $c > 0$ large enough, we achieve that

$$y_\infty > y^\diamond - \eta > y^\sharp.$$

From the N equation in (3.3) we obtain that

$$\limsup_{t\to\infty} \frac{1}{N}\frac{d}{dt}N \leq \beta - \mu - \alpha y_\infty < \beta - \mu - \alpha y^\sharp = 0,$$

by the definition of $y^\sharp$. Hence N decreases with a strictly negative exponential rate, a contradiction to our assumption that $C(N_\infty) > c > C(0)$.

Once it is established that the population size cannot tend to infinity, its boundedness can be easily derived from Theorem 1.3.

THEOREM 3.7. *Let $C(\infty) = \infty$, $\mathcal{P}_0 < 1$, and $y(0) > 0$. Then*

$$\limsup_{t\to\infty} N(t) < c < \infty,$$

with c not depending on the initial data.

Proof. We choose X as in the proof of Theorem 3.4 with the same metric such that X becomes a compact space. Set

$$\begin{aligned} X_1 &= \{(N,x,y,z); 0 \leq N < \infty, x,y,z \geq 0, x+y+z=1\}, \\ X_2 &= \{(\infty,x,y,z); x,y,z \geq 0, x+y+z=1\}. \end{aligned}$$

By Proposition 3.5, X_2 is a weak uniform repeller for X_1. As X and X_2 are compact, we can apply Theorem 1.3, and X_2 is a uniform strong repeller for X_1. This implies the statement of the theorem.

We mention that the condition $\mathcal{P}_0 < 1$ is almost sharp. For, if $\mathcal{P}_0 > 1$, the existence of a solution to (3.3) can be shown with N being exponentially increasing and y being bounded away from zero.

4. Uniform persistence via acyclicity of the boundary flow. As we have noticed in the previous section we cannot always obtain sharp conditions for uniform weak persistence by the fluctuation method. So we adapt an approach suggested by Butler and Waltman (1986), Butler, Freedman, and Waltman (1986), and Hale and Waltman (1989) for showing (uniformly) strong persistence.

Again we consider a continuous semiflow Φ on a metric space X, with X being the disjoint union of sets two X_1, X_2, where X_1 is open.

Note that we do neither assume that X_2 is contained in the closure of X_1 nor that X_2 is forward invariant.

4.1. Uniform weak persistence. As uniform weak persistence may be of interest in itself and requires less stringent compactness properties than uniform strong persistence, we discuss it in a section of its own.

We assume the following weaker version of the compactness assumption in §1:

($\mathbf{C}_{4.1}$) There is some $\delta > 0$ with the following properties:

- If $x \in X$ such that $d(\Phi_t(x), X_2) < \delta$ for all $t \geq 0$, then the forward orbit of x has compact closure in X.
- If x_n is a sequence in X satisfying

$$\limsup_{t\to\infty} d(\Phi_t(x_n), X_2) \to 0, \qquad n \to \infty,$$

then $\bigcup_{n\in\mathbf{N}} \omega(x_n)$ has compact closure.

The ω*-limit set* of a point y is defined as usual:

$$\omega(y) = \bigcap_{t\geq 0} \overline{\Phi([t,\infty) \times \{y\})}.$$

We say that an element $y \in X$ has a full orbit, if there is a function $x(t)$, $-\infty < t < \infty$, such that $x(0) = y$ and $x(t+s) = \Phi_t(x(s))$ for all $t \geq 0, s \in \mathbf{R}$. The α*-limit set* of a full orbit $x(t)$ is defined by

$$\alpha(x) = \bigcap_{t\geq 0} \overline{x((-\infty,-t])}.$$

We recall that a subset M of X is called *forward invariant* if and only if $\Phi_t(M) \subset M, t > 0$, and *invariant* if and only if $\Phi_t(M) = M, t > 0$. A compact invariant subset M of $Y \subseteq X$ is called an *isolated compact invariant set in* Y if there is an open subset

U of X such that there is no invariant set $\tilde{M}$ with $M \subseteq \tilde{M} \subseteq U \cap Y$ except M. U is called an *isolating neighborhood* of M.

PROPOSITION 4.1. *Let M be a compact invariant subset of X_2 which is an isolated compact invariant set in X. Let (x_n) be a sequence of elements in X such that*

$$\limsup_{t\to\infty} d(\Phi_t(x_n), X_2) \to 0, \qquad n \to \infty,$$

and $\omega(x_n) \not\subseteq M$. Further assume that there is a sequence $(p_n), p_n \in \omega(x_n)$, with $d(p_n, M) \to 0$, $n \to \infty$. Then, after choosing a subsequence, the sets $\omega(x_n)$ converge (in the Hausdorff metric) to a subset ω of X_2 that is invariant, compact, and connected. Further, there exist an element $u \in \omega \setminus M$ with $\omega(u) \subseteq M$ and an element $w \in \omega \setminus M$ with a full orbit in $\omega \subseteq X_2$ whose α-limit set is contained in M. u can be chosen such that its forward orbit is arbitrarily close to M. w can be chosen such that its backwards orbit is arbitrarily close to M.

Remark. If all elements x_n are identical, it is sufficient to assume that M is an isolated compact invariant set in X_2. This way we obtain a version of the Butler and McGehee lemma (Waltman (1991)).

We mention that the ω and α-limit sets in Proposition 4.1 are nonempty and compact because of the compactness condition $(\mathbf{C}_{4.1})$.

Proposition 4.1 should be compared to the almost identical Lemma 4.3 in Hale and Waltman (1989). Notice that we can arrange $u, w \in \omega \subseteq X_2$. Our proof parallels the proof of Lemma 4.3 in Hale and Waltman (1989). We give an even more detailed proof here in order to convince the reader that our compactness assumptions $\big((\mathbf{C}_{4.1})$ rather than point-dissipativity and asymptotic smoothness$\big)$ are sufficient.

Proof. In view of our first assumptions in $(\mathbf{C}_{4.1})$, we can assume that $d(\Phi_t(x_n), X_2) < \delta$ for all $t \geq 0$. Otherwise we replace x_n by $\Phi_{\tau_n}(x_n)$ with suitably chosen τ_n. It follows from our compactness condition $(\mathbf{C}_{4.1})$ that the closure Ω of $\bigcup_{n=1}^{\infty} \omega(x_n)$ is compact. Hence, after choosing a subsequence, the sets $\omega(x_n)$ converge towards a set ω in the Hausdorff metric. ω inherits compactness, invariance, and connectedness from the sets $\omega(x_n)$. Our assumptions further imply $\omega \subseteq X_2$.

Let U, V be open sets in X such that $M \subset V \subset \bar{V} \subset U$, and U is an isolating neighborhood of M. We can choose U as small as needed and in particular as a subset of the δ neighborhood of X_2. As U is isolating, we have $\omega(x_n) \not\subseteq U$ because otherwise U would contain the compact invariant set $\omega(x_n) \cup M$, which is larger than the maximal (in U) compact invariant set M. As $\omega(x_n) \ni p_n \to M$ we find sequences $r_n < s_n < t_n$ with

$$\Phi_{r_n}(x_n) \in \partial V, \quad \Phi_{t_n}(x_n) \in \partial V, \quad \Phi_{s_n}(x_n) \to M, \tag{4.1}$$

$$\Phi_t(x_n) \in V, \qquad r_n < t < t_n. \tag{4.2}$$

For each n, r_n, s_n, t_n can be chosen arbitrarily large. In particular we can arrange that

$$d(\Phi_{r_n}(x_n), \omega(x_n)) \to 0, \quad d(\Phi_{t_n - j}(x_n), \omega(x_n)) \to 0, \quad j = 1, \cdots, n, \quad n \to \infty. \tag{4.3}$$

As $\bigcup_{n=1}^{\infty} \omega(x_n)$ has compact closure, we can choose subsequences such that

$$\Phi_{r_n}(x_n) \to u \in \partial V \cap \omega, \quad \Phi_{s_n}(x_n) \to v \in M, \quad \Phi_{t_n}(x_n) \to w \in \partial V \cap \omega. \tag{4.4}$$

We claim $\omega(u) \subset M$ and that w has a full orbit whose α-limit set is contained in M.

In order to show $\omega(u) \subset M$, we first assume that the sequence $s_n - r_n$ is bounded. After choosing subsequences we may assume that $s_n - r_n \to \sigma$, $n \to \infty$. As Φ is continuous, it follows from (4.4) that

$$v = \lim_{n\to\infty} \Phi_{s_n}(x_n) = \lim_{n\to\infty} \Phi_{s_n - r_n}(\Phi_{r_n}(x_n)) = \Phi_\sigma(u).$$

As $v \in M$ and M is invariant, we have $\omega(u) = \omega(v) \subset M$.

Hence we can assume that, after choosing subsequences, the sequence $s_n - r_n$ converges to ∞ for $n \to \infty$. This implies that u has a full forward orbit in $\bar{V} \cap \omega$. As ω is compact, the ω-limit set of u is nonempty, compact, and invariant. Hence $\omega(u) \subset M$.

To show that w has a full orbit with α-limit set in M we proceed similarly. First we assume that the sequence $t_n - s_n$ is bounded. After choosing subsequences we may assume that $t_n - s_n \to \tau, n \to \infty$. As Φ is continuous, we have from (4.4) that

$$w = \lim_{n\to\infty} \Phi_{t_n - s_n}(\Phi_{s_n}(x_n)) = \Phi_\tau(v).$$

As M is invariant and $v \in M$, we have $w \in M$, a contradiction, because $w \in \partial V$. So we can assume that the sequence $t_n - s_n$ converges to ∞ after choosing subsequences again. Let $j \in \mathbf{N}$. Then $\Phi_{t_n - j}(x_n) \in V$ for sufficiently large n. Now recall (4.3) and (4.4) and the compactness of the closure of $\bigcup_{n=1}^\infty \omega(x_n)$. Employing a diagonalization procedure we can assume that, after choosing a subsequence,

$$\Phi_{t_n - j}(x_n) \to w_j \in \bar{V} \cap \omega, \qquad n \to \infty.$$

As Φ is continuous, we have $\Phi_j(w_j) = w$, $\Phi_1(w_j) = w_{j-1}$. The definition

$$\Phi_{-j+s}(w) = \Phi_s(w_j), \qquad 0 \le s < 1,$$

provides a continuous backward orbit in ω starting from w. Further, by (4.2),

$$\Phi_{-j+s}(w) = \lim_{n\to\infty} \Phi_s(\Phi_{t_n - j}(x_n)) = \lim_{n\to\infty} \Phi_{t_n + s - j}(x_n) \in \bar{V}.$$

Hence we have a full continuous backward orbit in $\bar{V} \cap \omega$ starting at w. This backward orbit has compact closure because ω is compact. Thus the α-limit set of this orbit of w is a nonempty, compact, invariant subset of the isolating neighborhood U, and hence a subset of M.

LEMMA 4.2. *Let x_n be a sequence of elements in X satisfying*

$$\limsup_{t\to\infty} d(\Phi_t(x_n), X_2) \to 0, \qquad n \to \infty.$$

Then, after choosing a subsequence, $\omega(x_n) \to \omega, n \to \infty$, in the Hausdorff metric, with some compact, invariant, connected subset ω of X_2. Moreover, for every

$$z \in \Omega_\omega := \bigcup_{y \in \omega} \omega(y),$$

there exists a sequence $p_n \in \omega(x_n)$ such that $p_n \to z, n \to \infty$.

We remark that the ω-limit sets in Lemma 4.2 are nonempty, compact, and invariant by our compactness assumption $(\mathbf{C}_{4.1})$.

Proof. We can assume that all forward orbits of x_n lie in the δ neighborhood of X_2. The existence of ω now follows as in Proposition 4.1. Let $z \in \Omega_\omega$. Then $z \in \omega(y)$ for some $y \in \omega$. Let $\epsilon > 0$. Then there is some $t > 0$ such that $d(\Phi_t(y), z) < \epsilon$. As Φ is continuous, there is some neighborhood W of y such that $d(\Phi_t(w), z) < \epsilon$ for any $w \in W$. As $\omega(x_n) \to \omega$ in the Hausdorff metric, we have $\omega(x_n) \cap W \neq \emptyset$ for all sufficiently large n. So, for any large enough n, we find some $q \in \omega(x_n)$ such that $d(\Phi_t(q), z) < \epsilon$. As $\omega(x_n)$ is invariant, $p = \Phi_t(q) \in \omega(x_n), d(p, z) < \epsilon$.

We define

$$\Omega_2 = \bigcup_{y \in Y_2} \omega(y), \qquad Y_2 = \{x \in X_2; \Phi_t(x) \in X_2 \; \forall t > 0\}.$$

Our compactness assumption $(\mathbf{C}_{4.1})$ implies that Ω_2 has compact closure and is invariant. Following Hale and Waltman (1989), a finite covering $M = \bigcup_{k=1}^m M_k$ in X_2 is called *isolated* if the sets M_k are pairwise disjoint subsets of X_2, which are isolated compact invariant sets in X.

A set $M \subset X_2$ is said to be *chained* (in X_2) to another (not necessarily different) set $N \subset X_2$, symbolically $M \mapsto N$, if there is some $y \in X_2, y \notin M \cup N$, and a full orbit through y in X_2 whose α-limit set is contained in M and whose ω-limit set is contained in N.

A finite covering $M = \bigcup_{k=1}^m M_k$ is called *cyclic* if, after possible renumbering, $M_1 \mapsto M_1$ or $M_1 \mapsto M_2 \mapsto \cdots \mapsto M_k \mapsto M_1$ for some $k \in \{2, \ldots, m\}$. M is called an *acyclic covering* otherwise.

Notice that Y_2 and so Ω_2 may be empty, e.g., if all orbits starting in X_2 leave X_2 and never return. Then Ω_2 has an acyclic covering, namely, the empty set. In many applications, however, X_2 is forward invariant such that $Y_2 = X_2$.

PROPOSITION 4.3. *Let x_n be a sequence of elements in X_1 satisfying*

$$\limsup_{t \to \infty} d(\Phi_t(x_n), X_2) \to 0, \qquad n \to \infty.$$

Let $M = \bigcup_{k=1}^m M_k$ be an isolated covering of Ω_2 such that $\omega(x_n) \not\subseteq M_k$ for all n, k. Then M is cyclic.

Proof. As $\omega \subseteq Y_2$, we obtain that $\Omega_\omega \subseteq \Omega_2$. By Lemma 4.2 (after possible renumbering of the covering M) we have $\Omega_\omega \cap M_1 \neq \emptyset$, and there is a sequence $p_n \in \omega(x_n)$ such that $d(p_n, M_1) \to 0, n \to \infty$. As $\omega(x_n) \not\subseteq M_1$, Proposition 4.1 provides an element $u \in \omega \setminus M_1$ with a full orbit in X_2 whose α-limit set lies in M_1. As $\omega \subset Y_2$, $\omega(u) \subset M$. As $\omega(u)$ is connected and the M_k are pairwise disjoint and compact, $\omega(u) \subset M_k$ for some k. Either $k = 1$ and the proof is finished or, after possible renumbering, $\omega(u) \subset M_2$ and $M_1 \mapsto M_2$. As $u \in \omega, \omega(u) \subset M_2$, we have that $\Omega_\omega \cap M_2 \neq \emptyset$, and Lemma 4.2 provides a (possibly different) sequence $p_n \in \omega(x_n)$ with $d(p_n, M_2) \to 0, n \to \infty$. Repeating our argument, we find that M_2 is chained to some M_k. Continuing this way we finally obtain a cyclic chain because there are only finitely many M_k.

THEOREM 4.4. *Let X_1 be open and forward invariant under the continuous semiflow Φ on X. Let the compactness assumption $(\mathbf{C}_{4.1})$ hold. Assume that Ω_2 has an acyclic isolated covering $M = \bigcup_{k=1}^m M_k$ such that each part M_k of M is a weak repeller for X_1. Then X_2 is a uniform weak repeller for X_1.*

Proof. If X_2 is not a uniform weak repeller for X_1, then, by definition (see the Introduction), we find a sequence $x_n \in X_1$ satisfying

$$\limsup_{t\to\infty} d(\Phi_t(x_n), X_2) \to 0, \qquad n \to \infty.$$

As each part M_k of M is a weak repeller for X_1, we have $\omega(x_n) \not\subseteq M_k$. Hence the assumptions of Proposition 4.3 are satisfied and the covering M has to be cyclic, in contradiction to the assumptions of the theorem.

4.2. Uniform strong persistence. The next theorem is general enough to obtain all results for our endemic model that can possibly be obtained from an acyclicity consideration.

THEOREM 4.5. *Let X be locally compact, and let X_2 be compact in X and X_1 be forward invariant under the continuous semiflow Φ on X. Assume that Ω_2,*

$$\Omega_2 = \bigcup_{y\in Y_2} \omega(y), \qquad Y_2 = \{x \in X_2; \Phi_t(x) \in X_2\ \forall t > 0\},$$

has an acyclic isolated covering $M = \bigcup_{k=1}^m M_k$. If each part M_k of M is a weak repeller for X_1, then X_2 is a uniform strong repeller for X_1.

In case that Y_2 is empty, Ω_2 has an acyclic isolated covering, the empty set.

Theorem 4.5 has been stated in Freedman and Moson (1990) as a remark in the following form: If the assumptions of Theorem 4.5 are satisfied and X_2 is forward invariant and a weak repeller for X_1, then it is a uniform strong repeller for X_1.

Proof. As X_2 has a compact neighborhood in X, the compactness assumption $(\mathbf{C}_{4.1})$ is automatically satisfied. Hence, by Theorem 4.4, X_2 is a uniform weak repeller for X_1. Thus X_2 is a uniform strong repeller for X_1 by Theorem 1.3.

If we face infinite-dimensional dynamical systems, more complicated assumptions have to be made. We require the following compactness condition rather than point dissipativity and asymptotic smoothness (Hale and Waltman (1989)):

$(\mathbf{C}_{4.2})$ There exist $\delta > 0$ and a subset B of X with the following properties:

- If $x \in X$ and $d(x, X_2) < \delta$, then $d(\Phi_t(x), B) \to 0, t \to \infty$.
- The intersection $B \cap B_\delta(X_2)$ of B with the δ-neighborhood of X_2, $B_\delta(X_2) = \{x \in X; d(x, X_2) < \delta\}$ has compact closure.

THEOREM 4.6. *Let X_1 be open in X and forward invariant under Φ. Further, let the compactness assumption $(\mathbf{C}_{4.2})$ hold. Assume that Ω_2,*

$$\Omega_2 = \bigcup_{y\in Y_2} \omega(y), \qquad Y_2 = \{x \in X_2; \Phi_t(x) \in X_2\ \forall t > 0\},$$

has an acyclic isolated covering $M = \bigcup_{k=1}^m M_k$ such that each part M_k of M is a weak repeller for X_1. Then X_2 is a uniform strong repeller for X_1.

Proof. Apparently the assumptions of Theorem 4.4 are satisfied such that X_2 is a uniform weak repeller for X_1. This implies that X_1 is forward invariant under Φ. We now refine the arguments in Proposition 1.2. As X_2 is a uniform weak repeller for X_1, there exists some $\epsilon > 0$ such that

$$\limsup_{t\to\infty} d(\Phi_t(x), X_2) > \epsilon \quad \forall x \in X_1. \tag{4.5}$$

We can assume that $\epsilon < \delta$ in assumption $(\mathbf{C}_{4.2})$.

Now suppose that X_2 is not a uniform strong repeller for X_1. Then there exist sequences $x_j \in X_1$ and $0 < \epsilon_j \to 0$ such that

$$\liminf_{t\to\infty} d(\Phi_t(x_j), X_2) < \epsilon_j \quad \forall j. \tag{4.6}$$

For every j we can find sequences $r_{jn} < s_{jn} < t_{jn}$ with $r_{jn} \to \infty, n \to \infty$, such that

$$d(\Phi_{s_{jn}}(x_j), X_2) < \epsilon_j, \tag{4.7}$$

$$d(\Phi_{r_{jn}}(x_j), X_2) = \epsilon = d(\Phi_{t_{jn}}(x_j), X_2), \tag{4.8}$$

$$d(\Phi_t(x_j), X_2) < \epsilon, \qquad r_{jn} < t < t_{jn}. \tag{4.9}$$

By (4.8) and assumption $(\mathbf{C}_{4.2})$ we have that $d(\Phi_{r_{jn}}(x_j), B \cap B_\delta(X_2)) \to 0, n \to \infty$. We choose numbers n_j such that $d(\Phi_{r_{jn}}(x_j), B \cap B_\delta(X_2)) < 1/j$. As $B \cap B_\delta(X_2)$ has compact closure,we can assume that, after choosing a subsequence, $(\Phi_{r_{jn_j}}(x_j))$ converges for $j \to \infty$. We set $r_j = r_{jn_j}, s_j = s_{jn_j}, t_j = t_{jn_j}$. By (4.7)–(4.9), we have $r_j < s_j < t_j$ with $r_j \to \infty, j \to \infty$ such that

$$d(\Phi_{s_j}(x_j), X_2) \to 0, \qquad j \to \infty, \tag{4.10}$$

$$d(\Phi_{r_j}(x_j), X_2) = \epsilon = d(\Phi_{t_j}(x_j), X_2), \tag{4.11}$$

$$d(\Phi_s(x_j), X_2) < \epsilon, \qquad r_j < s < t_j, \tag{4.12}$$

with $\Phi_{r_j}(x_j)$ having a limit for $j \to \infty$. Then we conclude as in Proposition 1.2. First we show that $t_j - r_j \to \infty$. If not, the sequence $\Phi_{s_j}(x_j)$ also converges, and we can arrive at a contradiction in the same way as in Step 1 of Proposition 1.2. To perform Step 2 in Proposition 1.2 we only need the convergence of $\Phi_{r_j}(x_j)$, which we have. This finishes the proof.

5. Application of the acyclicity method to an endemic model: persistence and limitation of the host population. (Act II).

5.1. Host persistence. We intend to derive a sharp condition for persistence of the host population in model (3.1) under the assumption $C(0) > 0$. We have to show that the set $X_2 = \{N = 0, x \geq 0, y \geq 0, z \geq 0, x+y+z = 1\}$ is a uniform strong repeller for $X_1 = \{N > 0, x \geq 0, y \geq 0, z \geq 0, x+y+z = 1\}$. We want to apply Theorem 4.5. To this end we analyze the semiflow induced by (3.3) on the forward invariant set X_2, i.e., for $N \equiv 0$:

$$\begin{aligned} \frac{d}{dt}x &= \beta(1-x) - (C(0)-\alpha)xy + \rho z, \\ \frac{d}{dt}y &= C(0)xy - (\gamma+\alpha+\beta)y + \alpha y^2, \\ \frac{d}{dt}z &= \gamma y - (\beta+\rho)z + \alpha yz, \\ 1 &= x+y+z. \end{aligned} \tag{5.1}$$

This is a special case of the model studied by Busenberg and van den Driessche (1990) and Busenberg and Hadeler (1990). They show the following.

PROPOSITION 5.1. (a) *Let* $C(0) \leq \beta+\alpha+\gamma$. *Then* $y(t), z(t) \to 0, x(t) \to 1$, $t \to \infty$.

(b) *Let* $C(0) > \beta+\alpha+\gamma$. *Then there exist unique equilibrium solutions to* (5.1) *such that* $x^\star, y^\star, z^\star > 0, x^\star+y^\star+z^\star = 1$. *Moreover,* $x(t), y(t)z(t) \to x^\star, y^\star, z^\star,$ $t \to \infty$, *for any solution x, y, z to* (5.1) *satisfying* $x(0) \geq 0, y(0) > 0, z(0) \geq 0, x(0) + y(0) + z(0) = 1$. *However, if x, y, z is a solution to* (5.1) *satisfying* $y(0) = 0,\ x(0) \geq 0, z(0) \geq 0,\ x(0) + z(0) = 1$, *then* $y \equiv 0$ *and* $x(t) \to 1, z(t) \to 0, t \to \infty$.

Hence $\Omega_2 = \bigcup_{\vec{y} \in X_2} \omega(\vec{y})$ consists of one or two elements, the disease-free equilibrium $N = 0, x = 1, y = 0 = z$, and the endemic equilibrium $N = 0, x = x^\star, y = y^\star, z = z^\star$. By Proposition 5.1 these equilibria cannot be chained to themselves. Furthermore, they cannot be chained to each other in a cyclic way. So they represent an acyclic covering for Ω_2. To show that this covering is isolated and a weak repeller for X_1, we analyze the behavior of

$$\frac{d}{dt}N = (\beta - \mu - \alpha y)N, \qquad N(0) > 0, \tag{5.2}$$

if y either stays close to zero or to $y^\star$ (in case that the latter exists). If y stays close to zero, N increases exponentially because we have assumed $\beta > \mu$. If we can show that N also increases exponentially if y stays close to $y^\star$, we are done: For every trajectory starting with $N(0) = 0$ converges to one of the two equilibria by Proposition 5.1, and no trajectory starting with $N(0) > 0$ can stay close to $N = 0, x = 1, y = z = 0$ or to $N = 0, x = x^\star, y = y^\star, z = z^\star$.

In our case it is actually possible to determine the equilibrium solutions in Proposition 5.1(b) explicitly. This seems no longer possible in the more complicated models by Busenberg and van den Driessche (1990) and Busenberg and Hadeler (1990). Even in our special case the explicit solution of $y^\star$ is not very helpful in analyzing (5.2) if y stays close to $y^\star$. So we try to obtain information from the equilibrium equations. At this point it is not clear whether our procedure would also work for the above-mentioned more complicated models. We can restrict our consideration to the case that $C(0) > \beta + \alpha + \gamma$:

$$\begin{aligned} 0 &= C(0)(1 - y^\star - z^\star) - (\gamma + \alpha + \beta) + \alpha y^\star, \\ 0 &= \gamma y^\star - (\beta + \rho)z^\star + \alpha y^\star z^\star. \end{aligned}$$

We solve both equations for $z^\star$:

$$\begin{aligned} z^\star &= 1 - y^\star - \frac{\gamma + \alpha + \beta}{C(0)} + \frac{\alpha}{C(0)} y^\star, \\ z^\star &= \frac{\gamma y^\star}{\beta + \rho - \alpha y^\star}. \end{aligned} \tag{5.3}$$

From the second equation we learn the constraint

$$\alpha y^\star < \beta + \rho; \tag{5.4}$$

otherwise $z^\star$ is not positive. We equate the two equations in (5.3) and reorganize terms:

$$0 = 1 - y^\star - \frac{\gamma + \alpha + \beta}{C(0)} + \frac{\alpha}{C(0)} y^\star - \frac{\gamma y^\star}{\beta + \rho - \alpha y^\star}. \tag{5.5}$$

In the range (5.4), the right-hand side of (5.5) is a strictly decreasing function of $y^\star$ (recall $C(0) > \gamma + \alpha + \beta > \alpha$). Hence there is at most one solution to (5.3). If $y^\star$ approaches the minimum of 1 and $(\beta + \rho)/\alpha$, the right-hand side of (5.5) becomes negative. Hence, by the intermediate value theorem, we have a nontrivial solution if and only if the right-hand side of (5.5) is strictly positive for $y^\star = 0$, i.e., if

$$1 - \frac{\gamma + \alpha + \beta}{C(0)} > 0.$$

We are actually interested in the question whether

$$\beta - \mu - \alpha y^\star > 0, \quad \text{i.e., } y^\star < \frac{\beta - \mu}{\alpha}.$$

(Then we can conclude that N in (5.2) increases exponentially if y stays sufficiently close to $y^\star$.) Assuming $C(0) > \gamma + \alpha + \beta$, this is the case if and only if the right-hand side of (5.5) is strictly negative for $y^\star = (\beta - \mu)/\alpha$, i.e.,

$$1 - \frac{\beta - \mu}{\alpha} - \frac{\gamma + \alpha + \mu}{C(0)} - \frac{\gamma(\beta - \mu)}{\alpha(\mu + \rho)} < 0.$$

We reorganize this inequality:

$$\frac{1}{C(0)} > \mathcal{T}_0$$

with

$$\mathcal{T}_0 := = \frac{1 - \mathcal{P}_0}{\mu + \gamma + \alpha.}, \qquad \mathcal{P}_0 = \frac{\beta - \mu}{\alpha}\left(1 + \frac{\gamma}{\mu + \rho}\right). \tag{5.6}$$

We combine these considerations to the following statement.

THEOREM 5.2. *Let*

$$\frac{1}{C(0)} > \mathcal{T}_0,$$

with $\mathcal{T}_0$ *being defined in* (5.6) *Then the host population is strongly uniformly persistent, i.e.,*

$$\liminf_{t \to \infty} N(t) \geq \epsilon > 0,$$

with ϵ *not depending on the initial data unless* $N(0) = 0$.

The result in Theorem 5.2 is almost sharp because it can be shown that there is a solution to (3.3) with $N(t)$ exponentially decreasing if $1/C(0) < \mathcal{T}_0$ (see Thieme, Thm. 7.1).

5.2. Host limitation. In §3.3 we have shown that the disease keeps the host population bounded if $C(\infty) = \infty$ and $\mathcal{P}_0 < 1$. Now we will derive a sharp condition for the case that $C(\infty) < \infty$. To this end we want to study an appropriate boundary flow. As the boundary of interest contains $N = \infty$, we first have to extend the state space and the dynamical system. This is done in the same way as in Theorem 3.4, and X becomes a compact metric space. We will show that the set $X_2 = \{(N, x, y, z); N, x, y, z \geq 0, x + y + z = 1, N = \infty \text{ or } y = 0\}$ is a uniform strong repeller for the set $X_1 = \{(N, x, y, z); 0 \leq N < \infty, x, z \geq 0, y > 0, x + y + z = 1\}$. We want to apply Theorem 4.5. To this end we analyze the flow on the forward invariant set X_2, i.e., $N \equiv \infty$ and

$$(5.7)\qquad \begin{aligned} \frac{d}{dt}x &= \beta(1-x) - (C(\infty)-\alpha)xy + \rho z, \\ \frac{d}{dt}y &= C(\infty)xy - (\gamma+\alpha+\beta)y + \alpha y^2, \\ \frac{d}{dt}z &= \gamma y - (\beta+\rho)z + \alpha yz, \\ 1 &= x+y+z. \end{aligned}$$

Again this is a special case of the model studied by Busenberg and van den Driessche (1990) and Busenberg and Hadeler (1990).

PROPOSITION 5.3. (a) *Let* $C(\infty) \leq \beta+\alpha+\gamma$. *Then* $y(t), z(t) \to 0, x(t) \to 1,\ t \to \infty$. (b) *Let* $C(\infty) > \beta+\alpha+\gamma$. *Then there exist unique equilibrium solutions to* (5.7) *such that* $x^\star, y^\star, z^\star > 0, x^\star+y^\star+z^\star = 1$. *Moreover,* $x(t), y(t)z(t) \to x^\star, y^\star, z^\star,\ t \to \infty$, *for any solution* x, y, z *to* (5.7) *satisfying* $x(0) \geq 0, y(0) > 0, z(0) \geq 0, x(0)+y(0)+z(0) = 1$. *However, if* x, y, z *is a solution to* (5.7) *satisfying* $y(0) = 0,\ x(0) \geq 0, z(0) \geq 0,\ x(0)+z(0) = 1$, *then* $y \equiv 0$ *and* $x(t) \to 1, z(t) \to 0, N(t) \to \infty$ *for* $t \to \infty$.

Apparently the disease can only limit the host population if it becomes endemic, i.e., in the situation described in Proposition 5.3(b). Thus we assume $C(\infty) > \beta+\alpha+\gamma$. Then the set $\Omega_2 = \bigcup_{\vec{y}\in X_2}\omega(\vec{y})$ consists of two equilibria, the disease-free equilibrium $N = \infty, x = 1, y = 0, z = 0$, and the endemic equilibrium $N = \infty, x = x^\star, y = y^\star, z = z^\star$. Apparently the disease-free and the endemic equilibrium cannot be chained to themselves or (in a cyclic way) to each other in X_2. So we only need to show that each of these two equilibria are isolated for the semiflow on X and that the two equilibria do not attract orbits that start in X_1. For the disease-free equilibrium, this follows from Proposition 5.3 and from Theorem 3.4, which states that y remains bounded away from zero if $y(0) > 0$ with the bound being independent of $y(0)$. To prove the same for the endemic equilibrium, we analyze the behavior of

$$\frac{d}{dt}N = (\beta - \mu - \alpha y)N$$

and derive a condition that guarantees that N decreases exponentially if the solution stays close enough to the endemic equilibrium. This will imply that this equilibrium is the largest forward invariant set in a sufficiently small neighborhood and cannot attract orbits starting in X_1. The exponential decrease of N will follow if

$$\eta - \mu - \alpha y^\star < 0, \quad \text{i.e., } y^\star > \frac{\beta-\mu}{\alpha},$$

for $y^\star, z^\star > 0$ satisfying

$$\begin{aligned} 0 &= C(\infty)(1-y^\star-z^\star) - (\gamma+\alpha+\beta) + \alpha y^\star, \\ 0 &= \gamma y^\star - (\beta+\rho)z^\star + \alpha y^\star z^\star. \end{aligned}$$

A consideration analogous to the one leading to Theorem 5.2 shows that N exponentially decreases if y, z approach $y^\star, z^\star$, provided that $\mathcal{T}_0 > 1/C(\infty)$. We observe that $\mathcal{T}_0$ in (5.6) is a strictly increasing function of μ as long as $\mu \leq \beta$. Thus

$$(5.8)\qquad \mathcal{T}_0 < \frac{1}{\beta+\gamma+\alpha}$$

such that $\mathcal{T}_0 > 1/C(\infty)$ implies the assumption in Proposition 5.3(b), $C(\infty) > \beta + \gamma + \alpha$.

We combine these considerations into the following result.

THEOREM 5.4. *Let $N(0) > 0, y(0) > 0$ and*

$$\mathcal{T}_0 > \frac{1}{C(\infty)} > 0.$$

Then

$$\limsup_{t\to\infty} N(t) \leq c < \infty, \qquad \liminf_{t\to\infty} y(t) \geq \epsilon > 0$$

with c, ϵ being independent of the initial data.

Notice that $\mathcal{T}_0 > 1/C(\infty)$ includes $\mathcal{P}_0 < 1$. This condition is almost sharp because it can easily be seen that there is a solution with y being bounded away from zero and N exponentially increasing if $\mathcal{T}_0 < 1/C(\infty)$ and $C(\infty) > \beta + \alpha + \gamma$.

Combining Theorem 5.2 and 5.4 we see that the population size is bounded away from zero and infinity if

$$\frac{1}{C(0)} > \mathcal{T}_0 > \frac{1}{C(\infty)}, \tag{5.9}$$

i.e., the disease limits the population growth without eradicating the population. As we will see later (Theorem 7.1), the disease is endemic under this condition irrespective of whether $C(\infty) < \infty$ or $C(\infty) = \infty$. For the interested reader we mention that (3.3) has an equilibrium solution under (5.9). We will show elsewhere (Thieme, to appear) that the above-mentioned special (equilibrium or exponential) solutions, if they exist, are locally asymptotically stable in an appropriate sense.

6. "Weak" implies "strong" without point dissipativity. In this section we make an attempt to drop the point-dissipativity assumption more or less completely in order to find a sharp condition for endemicity of the disease without requiring that the population size be bounded or that $C(\infty) < \infty$.

Assumptions 6.1. Though we do not want to assume point dissipativity, we would like to take advantage of the possibility that some components of the dynamical system finally become bounded. Hence we assume the following:

(A) There is a subset Y_1 of X_1 such that, for all $x \in X_1$, there is some time $t(x) > 0$ with $\Phi_t(x) \in Y_1$ for all $t \geq t(x)$.

As we cannot require that semiorbits are bounded, we assume that they satisfy some kind of conditional compactness.

($\mathbf{C}_{6.1}$) For any bounded subset B of X_1 and any $y \in Y_1$, the orbit $\Phi([0,\infty) \times \{y\}) \cap B$ has compact closure.

This set may be empty. Then it will be relatively compact automatically. ω-limit sets are defined as usual:

$$\omega(y) = \bigcap_{t\geq 0} \overline{\Phi([t,\infty) \times \{y\})}.$$

Under our assumptions, $\omega(y)$ may be empty or unbounded. ($\mathbf{C}_{6.1}$) guarantees that the intersection of $\omega(y)$ with any bounded closed set is compact. Moreover, we require that

($\mathbf{C}_{6.2}$) $\bigcup_{y\in Y_1} \omega(y) \cap B$ has compact closure for any bounded subset B of Y_1.

Dropping point dissipativity requires certain orbits to be repelled in a fixed period of time.

(R) For any sufficiently small $\epsilon > 0$ there is a subset D of X_1 and some δ, $0 < \delta < \epsilon$, and a time $t > 0$ with the following properties:

(i) There is no element $x \in Y_1 \setminus D$ such that $d(x, X_2) = \epsilon$ and $d(\Phi_s(x), X_2) < \epsilon$ for all $0 < s \leq t$.

(ii) If $x \in Y_1 \setminus D$ is an element in Y_1 with semiorbit in Y_1 and $r \in (0, t]$ is such that $d(x, X_2) = \epsilon = d(\Phi_r(x), X_2)$ and $d(\Phi_s(x), X_2) < \epsilon$ for all $0 < s < r$, then $d(\Phi_s(x), X_2) \geq \delta$ for all $0 \leq s \leq r$.

(iii) $D \cap Y_1 \cap S_\epsilon(X_2)$ is bounded, where $S_\epsilon(X_2) = \{x \in X; d(x, X_2) = \epsilon\}$ is the ϵ shell of X_2 in X.

THEOREM 6.2. *Let X be a metric space with metric d. Let X be the disjoint union of two sets X_1 and X_2, with X_1 open in X. Let Φ be a continuous semiflow on X_1 such that Assumptions* 6.1 *are satisfied.*

Then X_2 is a uniform strong repeller for X_1, whenever it is a uniform weak repeller for X_1.

Proof. We follow the lines of the proof of Proposition 1.2. Assume that X_2 is a uniform weak repeller for X_1. Then there exists $\epsilon > 0$ such that

$$\limsup_{t\to\infty} d(\Phi_t(x), X_2) > \epsilon \quad \forall x \in X_1. \tag{6.1}$$

We can assume that ϵ is sufficiently small for Assumption 6.1 **(R)** to apply.

Now suppose that X_2 is not a uniform strong repeller for X_1. Then there exist sequences $x_j \in X_1$ and $0 < \epsilon_j \to 0$ such that

$$\liminf_{t\to\infty} d(\Phi_t(x_j), X_2) < \epsilon_j \quad \forall j. \tag{6.2}$$

For every j we can find sequences $r_{jn} < s_{jn} < t_{jn}$ with $r_{jn} \to \infty, n \to \infty$, such that

$$d(\Phi_{s_{jn}}(x_j), X_2) < \epsilon_j, \tag{6.3}$$

$$d(\Phi_{r_{jn}}(x_j), X_2) = \epsilon = d(\Phi_{t_{jn}}(x_j), X_2), \tag{6.4}$$

$$d(\Phi_s(x_j), X_2) < \epsilon, \quad r_{jn} < s < t_{jn}. \tag{6.5}$$

Replacing x_j by $\Phi_{\tau_j}(x_j)$ with a suitably chosen τ_j, we can assume that $d(x_j, X_2) \leq \epsilon$ and the semiorbits of all x_j lie in Y_1.

We choose B, D and t, δ according to Assumption 6.1 **(R)**.

Case 1. For infinitely many j there are infinitely many r_{jn} such that $\Phi_{r_{jn}}(x_j) \in D$.

After choosing subsequences we can assume that $\Phi_{r_{jn}}(x_j) \in D$ for all j, n. Let $B_\epsilon = D \cap Y_1 \cap S_\epsilon(X_2)$. By Assumption 6.1 **(R)** (iii), B_ϵ is bounded. Further, by (6.4), $\Phi_{r_{jn}}(x_j) \in B_\epsilon$ for all j, n. In particular, the sequence $\Phi_{r_{jn}}(x_j)$ is bounded. By the compactness assumption ($\mathbf{C}_{6.1}$), we have that $\omega(x_j) \cap B_\epsilon$ is nonempty and compact, and $\Phi_{r_{jn}}(x_j) \to \omega(x_j) \cap B_\epsilon, n \to \infty$. We now choose numbers n_j such that $d(\Phi_{r_{jn_j}}(x_j), \omega(x_j) \cap B_\epsilon) < 1/j$. By the compactness assumption ($\mathbf{C}_{6.2}$), we can assume that, after choosing subsequences, $\Phi_{r_{jn_j}}(x_j)$ converges. Setting $r_j = r_{jn_j}, s_j = s_{jn_j}, t_j = t_{jn_j}$, we now find sequences $r_j < s_j < t_j$ with $r_j \to \infty, j \to \infty$, such that

$$d(\Phi_{s_j}(x_j), X_2) \to 0, \qquad j \to \infty, \tag{6.6}$$

$$d(\Phi_{r_j}(x_j), X_2) = \epsilon = d(\Phi_{t_j}(x_j), X_2), \tag{6.7}$$

$$d(\Phi_s(x_j), X_2) < \epsilon, \qquad r_j < s < t_j, \tag{6.8}$$

with $\Phi_{r_j}(x_j)$ having a limit for $j \to \infty$. Then we conclude as in Proposition 1.2. First we show that $t_j - r_j \to \infty$. If not, the sequences $\Phi_{s_j}(x_j), \Phi_{t_j}(x_j)$ also converge, and we can conclude as in Step 1 of Proposition 1.2. To perform Step 2 in Proposition 1.2 we only need the convergence of $\Phi_{r_j}(x_j)$, which we have.

Case 2. If Case 1 does not hold, we can find numbers $r_j < s_j < t_j$ with $r_j \to \infty, j \to \infty$, such that (6.6)–(6.8) hold and $\Phi_{r_j}(x_j) \in Y_1 \setminus D$.

We claim that $t_j - r_j \leq t$. If not, (6.7) and (6.8) contradict Assumption 6.1 **(R)** (i).

As $t_j - r_j \leq t$ and $\Phi_{r_j}(x_j) \in Y_1 \setminus D$, we conclude from (6.7), (6.8), and Assumption 6.1 **(R)** (ii) that $d(\Phi_s(x_j), X_2) \geq \delta > 0$ for all $r_j \leq s \leq t_j$. Noticing that δ does not depend on j, we realize that this contradicts (6.6).

7. Disease persistence in an endemic model. (Act III). Theorem 6.2 allows us to prove uniform strong persistence of the disease for model (3.3) also if $C(\infty) = \infty$. In Theorem 3.4 we have dealt with the case $\alpha + \beta + \gamma < C(\infty) < \infty$.

We want to show that $X_2 = \{(N, x, 0, z); 0 \leq N < \infty, x, z \geq 0, x + z = 1\}$ is a uniform strong repeller for the forward invariant set $X_1 = \{(N, x, y, z); 0 \leq N < \infty, x, z \geq 0, y > 0, x + y + z = 1\}$, if $C(\infty) > \alpha + \beta + \gamma$. Proposition 3.3 guarantees that X_2 is a uniform weak repeller for X_1. We check Assumptions 6.1. Note that

$$d((N, x, y, z), X_2) = y.$$

As for **(A)** we simply choose $Y = X$. $(\mathbf{C}_{6.1})$ and $(\mathbf{C}_{6.2})$ are satisfied because our state space lies in $\mathbf{R}^4$.

In order to determine D, δ, t (in dependence of ϵ) we make the following consideration.

Assume that $y(0) = \epsilon$, $y(s) \leq \epsilon$, $0 \leq s \leq t$. Notice from the N equation in (3.3) that N is exponentially increasing on [0,t] if ϵ is chosen small enough. Now, from the z equation in (3.3),

$$\frac{d}{dt}z \leq \gamma\epsilon - (\beta + \rho)z + \alpha\epsilon, \qquad 0 \leq s \leq t.$$

Here we have used that $z \leq 1$. Using this fact again we find

$$z(s) \leq e^{-(\beta+\rho)s} + \epsilon\frac{\gamma + \alpha}{\beta + \rho}, \qquad 0 \leq s \leq t.$$

Hence there is a constant $c > 0$ and some $s_\epsilon > 0$ such that

$$z(s) \leq c\epsilon, \qquad s_\epsilon \leq s \leq t.$$

We substitute (3.4) into the y equation in (3.3):

$$\frac{d}{dt}y = C(N)(1 - y - z)y - (\gamma + \alpha + \beta)y + \alpha y^2.$$

We obtain

$$\frac{\frac{d}{dt}y}{y} \geq \begin{cases} -(\gamma + \alpha + \beta), & 0 \leq s \leq s_\epsilon, \\ C(N)(1 - \epsilon[1 + c]) - (\gamma + \alpha + \beta), & s_\epsilon < s < t. \end{cases}$$

Let $\eta = C(\infty) - (\gamma + \alpha + \beta) > 0$. Let $N_0 > 0$ be such that

$$C(N_0) - (\gamma + \alpha + \beta) > \frac{2\eta}{3}.$$

Choose ϵ so small that

$$C(N_0)(1 - \epsilon[1 + c]) - (\gamma + \alpha + \beta) > \frac{\eta}{3}.$$

As $N(s) \geq N_0$ for $0 \leq s \leq t$, we have

$$\begin{aligned} y(s) &\geq \epsilon e^{-(\gamma+\alpha+\beta)s}, \qquad 0 \leq s \leq s_\epsilon, \\ y(s) &\geq \epsilon e^{-(\gamma+\alpha+\beta)s_\epsilon} e^{\eta(s-s_\epsilon)/3}, \qquad s_\epsilon \leq s \leq t. \end{aligned}$$

Choose $t = t_\epsilon$ such that

$$e^{-(\gamma+\alpha+\beta)s_\epsilon} e^{\eta(t_\epsilon - s_\epsilon)/3} > 1,$$

and $D = \{(N, x, y, z) \in X; N \leq N_0\}$. Revising the above considerations, we realize that **(R)**(i) is satisfied. **(R)**(ii) readily follows from the y equation in (3.3),

$$\frac{\frac{d}{dt}y}{y} \geq -(\gamma + \alpha + \beta),$$

and we can choose $\delta = \epsilon e^{-(\gamma+\alpha+\beta)t}$. **(R)**(iii) is obviously satisfied because D is bounded.

We now conclude the following from Theorem 6.2.

THEOREM 7.1. *Let $\alpha + \beta + \gamma < C(\infty)$. Then*

$$\liminf_{t\to\infty} y(t) \geq \epsilon > 0, \qquad \liminf_{t\to\infty} z(t) \geq \epsilon > 0,$$

with ϵ being independent of the initial data provided that $y(0) > 0$.

Proof. If $C(\infty) < \infty$, the statement follows from Theorem 3.4. If $C(\infty) = \infty$, the first estimate follows from X_2 being a uniform strong repeller for X_1. See the considerations at the beginning of this section. The second estimate is established as in the proof of Theorem 3.4 via the fluctuation method.

We mention that the condition in Theorem 7.1 is sharp because it can be shown that $y(t), z(t) \to 0$, $t \to \infty$, if $\alpha + \beta + \gamma \geq C(\infty)$ (see Thieme, Thm. 3.1).

Acknowledgment. Sections 1, 2, and 3 essentially developed when I taught a course on mathematical models for the spread of infectious diseases at Arizona State University in the fall of 1990. I thank my students Feng Zhilan and Liu Zhongmin for their interest and patience and then chairman W. T. Trotter for letting the course run in spite of the low enrollment. I thank Hal L. Smith and Paul Waltman for their valuable suggestions and comments.

REFERENCES

R. M. ANDERSON (1982), *Transmission dynamics and control of infectious diseases*, in Population Biology of Infectious Diseases, R. M. Anderson and R. M. May, eds., Life Sciences Research Report, 25, pp. 149–176; Dahlem conference, Springer-Verlag, Berlin, 1982.

R. M. ANDERSON AND R. M. MAY (1979), *Population biology of infectious diseases*: *Part* I, Nature, 280, pp. 361–367.

F. BRAUER (1990), *Models for the spread of universally fatal diseases*, J. Math. Biol., 28, pp. 451–462.

———(1991), *Models for the spread of universally fatal diseases,* II, in Differential Equations Models in Biology, Epidemiology and Ecology, S. Busenberg and M. Martelli, eds., pp. 57–69; Proceedings of the International Conference in Claremont, Jan. 1990, Lecture Notes in Biomath., 92, Springer-Verlag, New York.

S. BUSENBERG AND P. VAN DEN DRIESSCHE (1990), *Analysis of a disease transmission model in a population with varying size*, J. Math. Biol., 29, pp. 257–270.

S. BUSENBERG AND K. P. HADELER (1990), *Demography and epidemics*, Math. Biosci., 101, pp. 63–74.

G. BUTLER AND P. WALTMAN (1986), *Persistence in dynamical systems*, J. Differential Equations, 63, pp. 255–263.

G. BUTLER, H. I. FREEDMAN, AND P. WALTMAN (1986), *Uniformly persistent systems*, Proc. Amer. Math. Soc., 96, pp. 425–430.

C. CASTILLO-CHAVEZ, K. L. COOKE, W. HUANG, AND S. A. LEVIN (1989a), *On the role of long incubation periods in the dynamics of acquired immunodeficiency syndrome* (AIDS), *Part* 1. *Single population models*, J. Math. Biol., 27, pp. 373–398.

———(1989b), *On the role of long incubation periods in the dynamics of acquired immunodeficiency syndrome* (AIDS), *Part* 2. *Multiple group models*, in Mathematical and Statistical Approaches to AIDS-Epidemiology, C. Castillo-Chavez, ed., Lecture Notes in Biomath., 83, Springer-Verlag, pp. 177–189.

O. DIEKMANN AND M. KRETZSCHMAR (1991), *Pattern in the effects of infectious diseases on population growth*, J. Math. Biol., 29, pp. 539–570.

K. DIETZ (1982), *Overall population patterns in the transmission cycle of infectious disease agents*, in Population Biology of Infectious Diseases, R. M. Anderson and R. M. May, eds., Life Sciences Research Report 25; Dahlem conference, Springer-Verlag, Berlin, 1982, pp. 87–102.

A. FONDA (1988), *Uniformly persistent semidynamical systems*, Proc. Amer. Math. Soc., 104, pp. 111–116.

H. I. FREEDMAN AND P. MOSON (1990), *Persistence definitions and their connections*, Proc. Amer. Math. Soc., 109, pp. 1025–1032.

J. K. HALE AND P. WALTMAN (1989), *Persistence in infinite-dimensional systems*, SIAM J. Math. Anal., 20, pp. 388–395.

J. A. P. HEESTERBEEK AND J. A. J. METZ, *The saturating contact rate in marriage- and epidemic models*, J. Math. Biol., to appear.

W. M. HIRSCH, H. HANISCH, AND J. P. GABRIEL (1985), *Differential equation models for some parasitic infections*; *methods for the study of asymptotic behavior*, Comm. Pure Appl. Math., 38, pp. 733–753.

J. HOFBAUER AND K. SIGMUND (1988), *The Theory of Evolution and Dynamical Systems*, London Mathematical Society Student Texts 7, Cambridge University Press, London.

C. S. HOLLING (1966), *The functional response of invertebrate predators to prey density*, Mem. Ent. Soc. Canada, 48.

V. HUTSON AND K. SCHMITT (preprint), *Permanence in dynamical systems.*

A. PUGLIESE (1990), *Population models for diseases with no recovery*, J. Math. Biol., 28, pp. 65–82.

———(preprint), *An $S \longrightarrow E \longrightarrow I$ epidemic model with varying population size.*

H. R. THIEME (to appear), *Epidemic and demographic interaction in the spread of potentially fatal diseases in growing populations*, Math. Biosci.

H. R. THIEME AND C. CASTILLO-CHAVEZ (preprint), *How may infection-age dependent infectivity affect the dynamics of* HIV/AIDS?

P. WALTMAN (1991), *A brief survey of persistence in dynamical systems*, in Delay Differential Equations and Dynamical Systems, S. Busenberg and M. Martelli, eds., pp. 31-40. Proceedings of the International Conference in Claremont, Jan. 1990; Lecture Notes in Math. 1475, Springer-Verlag.

SIAM J. MATH. ANAL.
Vol. 24, No. 2, pp. 436–443, March 1993

A GEOMETRIC PROOF OF THE KWONG–MCLEOD UNIQUENESS RESULT*

C. B. CLEMONS† AND C. K. R. T. JONES‡

Abstract. Kwong recently proved the uniqueness of positive radial solutions of a semilinear elliptic equation with a superlinear term of a specific form. This was then generalized by Kevin McLeod to include a wide class of nonlinear terms. This paper gives a geometric context to the proof of this result. While these previous proofs used comparison functions that work apparently on account of certain striking calculations, this proof shows that the argument is equivalent to controlling a certain unstable manifold in a transformed phase space.

Key words. uniqueness, semilinear elliptic equation, Emden–Fowler transformation, unstable manifold

AMS(MOS) subject classifications. 35P30, 35P05, 34B15, 34F15

1. Introduction. In this paper we provide a geometric proof for the uniqueness of positive radially symmetric solutions of

$$\Delta u + f(u) = 0 \tag{1.1}$$

with appropriate boundary conditions and under certain assumptions about the nonlinear term $f(u)$, which was formulated by McLeod [6]. From the work of Gidas, Ni, and Nirenberg [2], we know that positive solutions of (1.1) on a ball are necessarily radially symmetric. Thus our uniqueness result yields uniqueness of positive solutions for the full problem (1.1), with appropriate boundary conditions.

Radially symmetric solutions of (1.1) satisfy the equation

$$u_{rr} + \frac{n-1}{r}u_r + f(u) = 0, \tag{1.2}$$

together with the boundary condition

$$u_r(0) = 0. \tag{1.3}$$

On a ball, of radius R, we further impose the Dirichlet boundary condition

$$u(R) = 0. \tag{1.4}$$

The first uniqueness result for positive solutions was proved by Coffman [1], in the case $n = 3$ and $f(u) = u^3 - u$. McLeod and Serrin [7] then generalized Coffman's result to obtain uniqueness for $f(u) = u^p - u$ and $1 < p < n/(n-2)$, provided $n > 2$, with further restrictions on this range when $n > 4$. Recently, Kwong [4] proved uniqueness in the model case $f(u) = u^p - u$ over the full range $1 < p < (n+2)/(n-2)$, thus giving uniqueness whenever there is existence (an existence proof due to Strauss can be found in [8]).

* Received by the editors July 11, 1991; accepted for publication (in revised form) June 8, 1992.

† Department of Mathematical Sciences, University of Akron, Akron, Ohio, 44325. The author was supported by National Science Foundation grant DMS-8801627.

‡ Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This author was supported by National Science Foundation grants DMS-8801627, DMS-9100085, and an award from the Graduate Research Board of the University of Maryland.

Kwong introduced the use of Sturm oscillation and a continuation argument in the space dimension n. Recently, McLeod [6] greatly simplified Kwong's argument while generalizing the allowable $f(u)$ and avoiding the continuation argument. Simultaneous to McLeod's proof, Kwong and Zhang obtained another general result [5]. Our proof is motivated by that of McLeod.

Both Kwong and McLeod compared the roots of $\delta u = \delta u(r, \alpha)/\delta\alpha$ with the function $v := \lambda u + r\dot{u}$. The quantity δu satisfies the equation of variations for (1.2), and the determination of its zeros naturally leads to uniqueness results; however, the function v, which is used to control it, appears to work on account of certain striking calculations. These calculations lead to an equation for v which is a forced version of the equation of variations in which the forcing term depends only on u and λ. It turns out that the comparison function used by both Kwong and McLeod is related to the vector field of a transformed version of (1.2).

We shall cast (1.1) as a first-order system and study the resulting phase space from a geometric point of view. In order to make an autonomous system, we introduce r as a dependent variable and change the independent variable to t via $r = e^t$; thus (1.2) becomes the system

$$\begin{aligned} \dot{u} &= rv, \\ \dot{v} &= -(n-1)v - rf(u), \qquad \dot{x} = \frac{d}{dt} \\ \dot{r} &= r. \end{aligned} \tag{1.5}$$

Note that the singularity $r = 0$ has now been removed and replaced by an invariant plane! The set of solutions satisfying the boundary condition (1.3) form a manifold in the phase space of (1.5), namely, the unstable manifold of the line of critical points $\{v = r = 0\}$.

Uniqueness can also be approached through a study of the geometric contortions of this manifold. This in turn is studied using tangent vectors to the manifold that can be followed along trajectories as solutions of the equations of variation. The natural tangent vectors are the vector field itself and a vector tangent to the intersection of the manifold with a fixed r-plane. Indeed the approach described above for proving uniqueness amounts to a comparison between these two vectors. The comparison of these vectors is interpreted here as control over the tilting of this manifold. This control can also be achieved through determination of the normal to this unstable manifold, which is formed by taking the cross product of the vector field with this tangent vector. In fact, for uniqueness it turns out that we need to determine the sign of an appropriate component of this normal.

Unfortunately, this comparison (or equivalently control over the normal) does not work in (1.5) due to the changes of sign of both $f'(u)$ and the first component of the tangent vector; however, after a change of variables, this strategy can be rescued. The change of variables is known as the Emden–Fowler transformation; see [3], where $y = r^\lambda u$, whence (1.5) becomes

$$\begin{aligned} \dot{y} &= \nu, \\ \dot{\nu} &= -(-2\lambda + n - 2)\nu - \lambda(\lambda - n + 2)y - r^{\lambda+2} f(r^{-\lambda} y), \\ \dot{r} &= r. \end{aligned} \tag{1.6}$$

In these variables, the differential equation which the normal satisfies involves the same expression that appears as a forcing term in McLeod's approach. As mentioned above, a comparison between a tangent vector and a vector field is related to control over a normal. It is not evident in the Kwong–McLeod approach that the comparison is between a tangent vector and a vector field. However in Emden–Fowler variables the comparison between tangent vector and vector field is equivalent to the argument they use.

Our goal in this paper is then to give an alternative proof to McLeod's result, which is naturally motivated by the geometry and we believe explains the underlying mechanism for uniqueness.

The following assumptions will be imposed on the nonlinear term $f(u)$:

(1) $f(u)$ is in $C^1([0,\infty))$;

(2) $f(0) = 0$;

(3) There is a β so that $f(\beta) = 0$ and $f(u) < 0$ if $0 < u < \beta$.

These are standard conditions which say the zeros of $f(u)$ are like those of $u^p - u$. The most unusual condition assumed on $f(u)$ was formulated by McLeod and it is this condition that permits control over the forcing term mentioned above.

(4) Let $I(u,\lambda) = (\lambda + 2)f(u) - \lambda u(df/du)(u)$. Assume that for each $U > \beta$ there is a $\lambda = \lambda(U) > 0$ such that

$$I(u,\lambda) \leq 0 \quad \text{for } 0 \leq u < U, \quad \text{and}$$

$$I(u,\lambda) \geq 0 \quad \text{for } u > U.$$

THEOREM. *If $f(u)$ satisfies the above conditions, then* (1.2) *with boundary conditions* (1.3), (1.4) *have at most one positive solution.*

Remarks. (1) It is a pleasant exercise to check that $f(u) = u^P - u$ satisfies condition (5), and that as $\lambda \to 2/(p-1)$, $U \to \infty$.

(2) The assumption imposed by McLeod that λ depends continuously on U is not needed in our proof.

(3) The λ in (5) will be realized as the exponent in Emden–Fowler transformation.

2. Set-up and basic lemmas. Consider again (1.5). The $\{r = 0\}$-plane is invariant, and its flow is governed by

$$\begin{aligned} \dot{u} &= 0 \\ \dot{v} &= -(n-1)v \end{aligned} \qquad \cdot = \frac{d}{dt}. \tag{2.1}$$

Note that the u-axis is a line of stable points and that each vertical line is invariant. A natural phase space for problem (1.5) is $R^2 \times [0,\infty)$. The boundary conditions translate to looking for a trajectory $(u(t),\ v(t),\ r(t))$ that satisfies

$$\lim_{t\to-\infty}(u(t),\ v(t),\ r(t)) = (\alpha,0,0), \quad \text{and}$$

$$\lim_{t\to\tau}(u(t),\ v(t),\ r(t)) = (0,\beta,R), \quad \tau = \ell n R,$$

where $\beta < 0$ due to the flow and the following phase portrait development.

At a critical point in the $\{r = 0\}$ plane, the linearization is

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1-n & f(\alpha) \\ 0 & 0 & 1 \end{pmatrix}$$

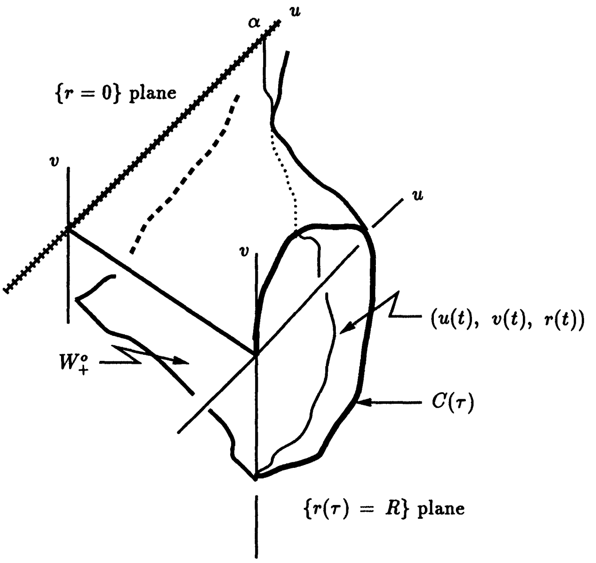


FIG. 1

with eigenvalues 0, $1 - n$, and 1. The stable eigenvector is (0,1,0). Eigenvectors associated with 0 and 1 are (1,0,0) and $(0,\ -f(\alpha)/n,\ 1)$, respectively. The unstable manifold points in the $r > 0$ direction. Near α, the local unstable manifolds foliate the center-unstable manifold of $(\alpha, 0, 0)$ W^{cu}_{loc}. Now set $W^o_{\text{loc}} : = \cup_{\alpha>0} W^u_{\text{loc}}(\alpha)$; then $W^o_+ : = \cup_{t>0}(W^o_{\text{loc}} \cdot t)$ is the surface containing all the solutions to problem (1.6).

The manifold W^o_+ is examined in more detail so as to isolate the properties associated with uniqueness. A solution $u(r)$ to problem (1.2) would satisfy $\dot{u}(0) = 0$, $u(R) = 0$, and in light of the last section, $u(0) = \alpha$ where α is a positive constant; henceforth, denote $u(r) : = u(r, \alpha)$. Similarly we denote a solution $(u(t),\ v(t),\ r(t))$ to problem (1.5) contained in W^o_+ by $u(t) := u(t, \alpha)$, $v(t) := v(t, \alpha)$, and $r(t)$, where α is given by $\lim_{t \to -\infty} u(t) = \alpha$.

The presence of more than one solution to problem (1.2)–(1.4) would say that there exists two trajectories, $(u(t, \alpha_1),\ v(t, \alpha_1),\ r(t))$ and $(u(t, \alpha_2), v(t, \alpha_2),\ r(t))$ of (1.5), contained in W^o_+ such that $(u(\tau, \alpha_1),\ v(\tau, \alpha_1),\ r(\tau)) = (0, \beta_1, R)$ and $(u(\tau, \alpha_2),$ $v(\tau, \alpha_2) r(\tau)) = (0, \beta_2, R)$.

Consider the intersection of W^o_+ with the $\{r(\tau) = R\}$-plane, $C(\tau) : = W^o_+ \bigcap \{\tau\}$; see Fig. 1. Nonuniqueness says that $C(\tau) \cap \{u = 0\}$ contains at least two elements. Center-unstable manifold theory gives that W^o_+ is sufficiently smooth so that the following may be defined. Let $T(t) = (\delta u(t, \alpha),\ \delta v(t, \alpha))$ be a tangent vector to $C(\tau)$ at the point $(u(t, \alpha),\ v(t, \alpha),\ r(t))$. When δr is set equal to zero, $(\delta u(t, \alpha), \delta v(t, \alpha))$ is a solution to the equation of variations associated with system (1.5),

$$\begin{aligned} \dot{\delta u} &= r \delta v, \\ \dot{\delta v} &= -(n-1)\delta v - r \frac{df(u)}{du} \delta u. \end{aligned} \tag{2.2}$$

Notice that $\lim_{t \to -\infty} \delta u(t, \alpha) = 1$, $\lim_{t \to -\infty} \delta v(t, \alpha) = 0$.

Two lemmas are used to prove uniqueness for $r(t) = R < \infty$. The first lemma sets up the necessary configuration of $C(t)$ in the case of nonuniqueness. This is related to a concept defined by Kwong as admissibility that is not strict. The second lemma argues that this configuration is not possible.

LEMMA 1. *Let τ be such that $r(\tau) = R$. Nonuniqueness to problem* (1.1)–(1.3) *implies that there exists a trajectory $(u(t,\hat{\alpha}),\ v(t,\hat{\alpha}),\ r(t))$ contained in W^o_+ with the following properties: there exist τ and $\hat{\alpha}$, where $-\infty < \tau_o < \hat{\tau} < \tau$, such that $u(\hat{\tau},\hat{\alpha}) = 0, \delta u(\tau_o,\hat{\alpha}) = 0$, and $\delta u(t,\hat{\alpha}) \neq 0$ for $t \in (-\infty,\tau_o) \cup (\tau_o,\hat{\tau})$. Moreover, $\delta u(\hat{\tau},\hat{\alpha}) = 0$, and $\delta v(\hat{\tau},\hat{\alpha}) < 0$.*

Proof of Lemma 1. Nonuniqueness at τ implies that there exist two trajectories, $(u(t,\alpha_1),\ v(t,\alpha_1),\ r(t))$ and $(u(t,\alpha_2),\ v(t,\alpha_2),\ r(t))$, contained in W^o_+ such that $u(\tau,\alpha_1) = u(\tau,\alpha_2) = 0$. Suppose $\alpha_1 < \alpha_2$ and, for no other trajectory in W^o_+ with $\alpha \in (0,\alpha_1) \cup (\alpha_1,\alpha_2)$, does $u(\tau,\alpha) = 0$; see Fig. 2.

Via continuity with respect to initial conditions and the intermediate value theorem, there exist $\hat{\alpha}$ and $\hat{\tau}$, where $\alpha_1 < \hat{\alpha} < \alpha_2$ and $\hat{\tau} < \tau$, such that $C(\hat{\tau})$ is tangent to the v-axis, and $u(\hat{\tau},\hat{\alpha}) = \delta u(\hat{\tau},\hat{\alpha}) = 0$.

Let $C(t,\alpha)$ be the portion of the curve $C(t)$ defined by $C(t,\alpha) = \{(u(t,\alpha), v(t,\alpha), r(t)) \in C(t) | a \in (0,\alpha]\}$. Also define the following curves in the tangent bundle to W^0_+ : $S_{C(\hat{\tau},\hat{\alpha})} = \{(\delta u(\hat{\tau},\alpha), \delta v(\tau,\hat{\alpha}),\ 0)$ for $(u(\hat{\tau},\alpha),\ v(\hat{\tau},\alpha),\ r(\tau)) \in C(\hat{\tau},\hat{\alpha})\}$ and $S_\gamma = \{(\delta u(t,\hat{\alpha}),\ \delta v(t,\hat{\alpha}))$ for $(u(t,\hat{\alpha}), v(t,\hat{\alpha}),\ r(t)) \in \gamma\}$, where γ is the curve $\gamma = \{(u(t,\hat{\alpha}),\ v(t,\hat{\alpha}),\ r(t)) | t \in (-\infty,\hat{\tau}]\}$. Observe that $C(\hat{\tau},\hat{\alpha})$ is homotopic to γ because they bound a portion of W^o_+ together with the u-axis and the r-axis. Also, $S_{C(\hat{\tau},\hat{\alpha})}$ is homotopic to S_γ.

Let I denote the winding number of a curve as in [3]. Since the winding number is homotopy invariant and $\delta\dot{u} = r\delta v$, we have $I(S_\gamma)=$ the number of zeros of $\delta u(t,\hat{\alpha})$ for $-\infty < t < \hat{\tau}$. Now, either $v(\tau,\alpha_1) > v(\tau,\alpha_2)$ or $v(\tau,\alpha_2) > v(\tau,\alpha_1)$. The first situation would imply that $I(S_\gamma) = I(C(\hat{\tau},\hat{\alpha})) = 1$. Because $(u(t),\ v(t)) = (\beta,0)$ is a solution of (4.5) and is contained in W^o_+, the first root of $\delta u(t,\hat{\alpha})$ must occur at some time before $u(t,\hat{\alpha}) = \beta$; see Fig. 2. If $v(\tau,\alpha_1) > v(\tau,\alpha_2)$, then $I(S_\gamma) = I(C(\hat{t},\hat{\alpha})) = 2$, and hence $\delta u(t,\hat{\alpha})$ has one root for $t \in (-\infty,\hat{\tau})$, thus proving the lemma.

Now consider (1.6), which results from applying the Emden–Fowler transformation. The origin is a critical point. Linearizing about this point produces the matrix

$$\begin{pmatrix} 0 & 1 & 0 \\ -\lambda(\lambda+2-n) & -(n-2-\lambda) & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The eigenvalues of the matrix are $\lambda, \lambda-n+2$, and 1. The associated eigenvectors are $(1,\lambda,0)$, $(1,\lambda-n+2,0)$, and $(0,0,1)$, respectively. If $\lambda < n-2$, a two-dimensional (or three-dimensional if $\lambda > n-2$) unstable manifold W^u_{loc} is associated with the origin. Set $W^u = \cap_{t>0} W^u_{\text{loc}} \cdot t$. The transformation from (u,v,r) to (y,ν,r), defined on $R^2 \times (0,\infty)$ by

$$T: \begin{cases} y = r^\lambda u, \\ \nu = r^{\lambda+1}v - \lambda r^\lambda u, \\ r = r, \end{cases} \tag{2.3}$$

carries W^u_{loc} to a submanifold of W^u, call is W^u_{pos}. Next consider the curve $C(\hat{\tau})$ defined in the proof of Lemma 1. Let $D(\hat{\tau}) = T(C(\hat{\tau}))$; then $D(\hat{\tau})$ is a curve in the $\{r = \hat{r}\}$ plane of the phase space associated with (1.7). Using the obvious definition

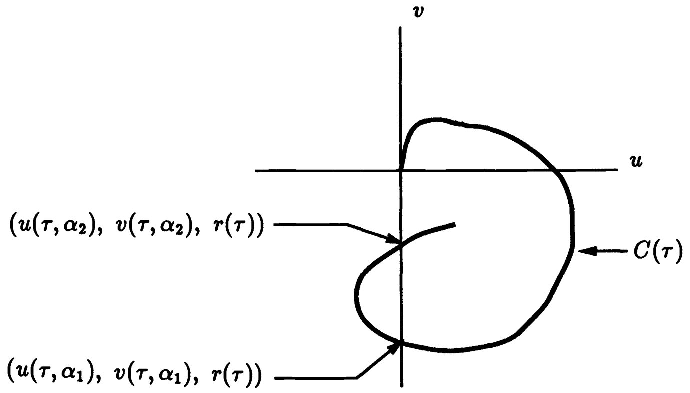

v
u
$(u(\tau,\alpha_2),\ v(\tau,\alpha_2),\ r(\tau))$
$C(\tau)$
$(u(\tau,\alpha_1),\ v(\tau,\alpha_1),\ r(\tau))$

FIG. 2

of the variables

$$\delta y(t,\alpha)=\left.\frac{\partial y(t,a)}{\partial \alpha}\right|_{a=\alpha},\qquad \delta v(t,\alpha)=\left.\frac{\partial v(t,a)}{\partial a}\right|_{a=\alpha},$$

we have the exact analogue of Lemma 1 in these variables; see Fig. 3.

LEMMA 2. *With the same hypothesis as Lemma* 1 *and $\tau = \ell nR$, there exists a trajectory $(u(t,\hat{\alpha}),\ v(t,\hat{\alpha}),\ r(t))$ contained in W^u_{pos} with the following properties: there exist τ_o and $\hat{\alpha}$, where $-\infty < \tau_o < \hat{\tau} < \tau$, such that $y(\hat{\tau},\hat{\alpha}) = 0$, $\delta y(\tau_o,\hat{\alpha}) = 0$, and $\delta y(t,\hat{\alpha}) \neq 0$ for $t \in (-\infty,\tau_o) \cup (\tau_o,\hat{\tau})$. Moreover, $\delta v(\hat{\tau},\hat{\alpha}) < 0$.*

Proof of Lemma 2. Lemma 2 follows easily from Lemma 1 because the zeros of $u(t,\hat{\alpha})$ and $y(t,\hat{\alpha})$ agree since $y = r^{\lambda}u$. Furthermore, $\delta y(t,\hat{\alpha}) = r^{\lambda}\delta u(t,\hat{\alpha})$, and so the zeros of δy and δu agree. The fact that $\delta v(\hat{\tau},\hat{\alpha}) < 0$ follows easily from the expression for δv obtained by differentiating (2.2).

3. Uniqueness proof. We shall interpret Lemma 2 in terms of the normal to the manifold W^u_{pos}. This is natural since Lemma 2 is a statement about a configuration of the unstable manifold that is forced by the assumption of nonuniqueness.

The vector field $(\dot{y},\dot{\nu},\dot{r})$ and the tangent vector $(\delta y,\delta\nu,0)$ are both tangent to W^u_{pos}. We denote $N(t,\alpha)$, the cross product of these two vectors, (in the order given) and use dual notation for its components.

$$N(t,\alpha)=(\delta y^*(t,\alpha),\ \delta\nu^*(t,\alpha),\ \delta r^*(t,\alpha)).$$

We then have the following.

LEMMA 3. *On the trajectory of Lemma* 2, *we have that the third component of the normal, $\delta r^*(\hat{\tau},\hat{\alpha}) > 0$.*

Proof.

$$\delta r^* = \begin{vmatrix} \dot{y} & \dot{\nu} \\ \delta y & \delta\nu \end{vmatrix} = \dot{y}\delta\nu - \dot{\nu}\delta y \quad \text{and} \quad t=\hat{\tau}, \alpha=\hat{\alpha}\delta y = 0, \delta\nu < 0 \quad \text{and} \quad \dot{y} < 0.$$

An equation for the components of the normal can be easily calculated. The

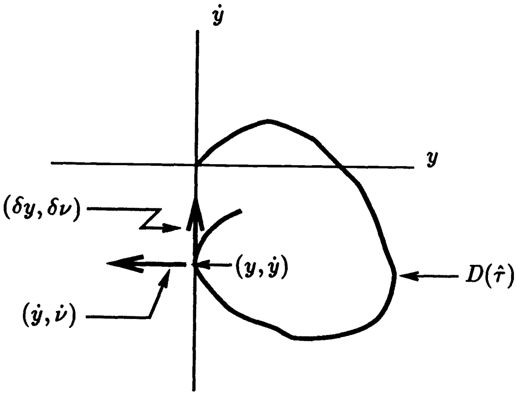


FIG. 3

equation of variations of (1.6) is

$$\delta\dot{y} = \delta\nu,$$

$$\delta\dot{\nu} = -(-2\lambda + n - 2)\delta\nu - \lambda(\lambda - n + 2)\delta y - r^2 \frac{df}{du}\delta y \tag{3.1}$$

$$- \frac{\partial}{\partial w}\left[r^{\lambda+2} f(r^{-\lambda} y)\right] \delta r,$$

$$\delta\dot{r} = \delta r.$$

If we abbreviate this as

$$\begin{pmatrix} \delta y \\ \delta\nu \\ \delta r \end{pmatrix}^{\cdot} = A \begin{pmatrix} \delta y \\ \delta\nu \\ \delta r \end{pmatrix}, \tag{3.2}$$

we can write the equation of the normal as

$$\begin{pmatrix} \delta y^* \\ \delta\nu^* \\ \delta r^* \end{pmatrix}^{\cdot} = (-A^* + (Tr\ A)I) \begin{pmatrix} \delta y^* \\ \delta\nu^* \\ \delta r^* \end{pmatrix}, \tag{3.3}$$

where A^* is the transpose of A. Using the identity

$$-A^*(\eta_1 \times \eta_2) = A\eta_1 \times \eta_2 + \eta_1 \times A\eta_2 - TrA(\eta_1 \times \eta_2),$$

from (3.3) we can calculate

$$(\delta r^*)^{\cdot} = +\frac{\partial}{\partial r}\left[r^{\lambda+2} f(r^{-\lambda} y)\right] \delta y^* - (-2\lambda + n - 2)\delta r^*. \tag{3.4}$$

Now

$$\frac{\partial}{\partial r}\left[r^{\lambda+2} f(r^{-\lambda} y)\right] = r^{\lambda+1}\left[(\lambda+2) f(u) - \lambda u \frac{df}{du}\right] = r^{\lambda+1} I(u, \lambda).$$

Moreover,

$$\delta y^* = -\begin{vmatrix} \dot{y} & \dot{r} \\ \delta y & 0 \end{vmatrix} = r\delta y,$$

so that (3.4) becomes

$$(\delta r^*)^{\cdot} = +r^{\lambda+2}I(u,\lambda)\delta y - (-2\lambda + n - 2)\delta r^*. \tag{3.5}$$

Using equation (3.5) and Lemma 3 we can complete the proof of the Theorem. The parameter λ is still free to be chosen. The quantity $\delta y(t, \hat{\alpha})$ changes sign at a certain value of $t = \tau_o$, and this is independent of λ. Let $U(\tau_o, \hat{\alpha}) = \hat{u}$. Then since u is monotone along the trajectory assumed by Lemma 2 we can force $I(u, \lambda)$ to change sign exactly once, at $\hat{u}$. We then see that $I(u, \lambda)\delta y$ has a fixed sign and it is easy to check that $I(u, \lambda)\delta y < 0$. We check that $e^{-2\lambda+(n-2)t}\delta r^* \to 0$ as $t \to -\infty$. But then $\delta r^*(\hat{\tau}, \hat{\alpha}) < 0$ from (3.5), which contradicts Lemma 3. It follows that the configuration described above is not possible, and hence uniqueness holds.

REFERENCES

[1] C. V. Coffman, *Uniqueness of the ground state solutions for* $\Delta u + u - u^3 = 0$ *and a variational characterization of other solutions*, Arch. Rational Mech. Anal., 46 (1972), pp. 81–95.

[2] B. Gidas, N. M. Ni, and L. Nirenberg, *Symmetry of positive solutions of nonlinear elliptic equations in* $\mathbf{R}^n$, Adv. Math. Studies, 7A (1981), pp. 369–402.

[3] C. K. R. T. Jones and T. Küpper, *On the infinitely many solutions of a semilinear elliptic equation*, SIAM J. Math. Anal., 17 (1986), pp. 803–835.

[4] M. K. Kwong, *Uniqueness of positive solutions of* $\Delta u - u + u^p = 0$ *in* $\mathbf{R}^n$, Arch. Rational Mech. Anal., 105 (1989), pp. 243–266.

[5] M. K. Kwong and L. Zhang, *Uniqueness of the positive solution of* $\Delta u + f(u) = 0$ *in an annulus*, Differential Integral Equations, 4 (1991), pp. 583–599.

[6] K. McLeod, *Uniqueness of positive solutions of* $\Delta u + f(u) = 0$ *in* $\mathbf{R}^n$, Tech. Report Series of the Department of Mathematical Sciences 3, University of Wisconsin-Milwaukee, 1989.

[7] K. McLeod and J. Serrin, *Uniqueness of positive radial solutions of* $\Delta u + f(u) = 0$ *in* $\mathbf{R}^n$, Arch. Rational Mech. Anal., 99 (1987), pp. 115–145.

[8] W. A. Strauss, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys., 55 (1977), pp. 149–162.

SIAM J. MATH. ANAL.
Vol. 24, No. 2, pp. 444–465, March 1993

011

GEOMETRIC ANALYSIS OF A NONLINEAR BOUNDARY VALUE PROBLEM FROM PHYSICAL OCEANOGRAPHY*

STEVEN R. DUNBAR†

Abstract. A third-order nonlinear differential equation with two sets of boundary conditions is considered. These boundary value problems arise as boundary layer problems from a model of large scale ocean circulation. Using geometrical techniques from qualitative differential equations, such as Wazewski's theorem, invariant manifolds, and Lyapunov functions, the existence of solutions for each boundary value problem is given in a uniform way for all positive values of a parameter of the differential equation.

Key words. nonlinear boundary layer problem, Wazewski's theorem, Lyapunov function, La Salle invariance principle

AMS(MOS) subject classifications. 34B15, 34C30, 34C35, 86A05

1. Introduction.

This paper presents a geometric analysis of the equation

$$\phi''' + \lambda(\phi\phi'' - (\phi')^2) + 1 - \phi = 0 \tag{1}$$

with the "no-slip" boundary conditions

$$\begin{aligned} \phi(0) &= 0, \\ \phi'(0) &= 0, \\ \phi(\infty) &= 1 \end{aligned} \tag{2}$$

and also with the "stress-free" boundary conditions

$$\begin{aligned} \phi(0) &= 0, \\ \phi''(0) &= 0, \\ \phi(\infty) &= 1 \end{aligned} \tag{3}$$

for all positive values of the parameter λ. This equation and its boundary values arise from a similarity solution of the so-called barotropic quasi-geostrophic potential vorticity equation for one layer ocean circulation. A derivation of the equation and further references to the oceanography literature are in [11].

In [11] Ierley and Ruehr present extensive numerical calculations and matched asymptotic expansions for the solutions of (1)–(2) and (1)–(3) in several parameter cases. For (1)–(2) Ierley and Ruehr set $\phi''(0) = \beta$ and then numerically determined values β for each parameter value λ so that the boundary value problem has a solution. Thus $\beta(\lambda)$ uniquely identifies a solution to the boundary value problem through the initial conditions $\phi(0) = 0$, $\phi'(0) = 0$, $\phi''(0) = \beta$. In Fig. 1 of [11] there is a curve in the λ-β plane parametrically representing the solutions. For $\lambda \geq 0$ there is only a

*Received by the editors September 6, 1989; accepted for publication (in revised form) May 26, 1992. This research was supported in part by National Science Foundation grant MCS-8601528.

†Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska, 68588.

single solution, corresponding to a value of $\beta \approx 1$. For λ in the range $\lambda_c \leq \lambda < 0$, where $\lambda_c \approx -0.79130$, there are two solutions which coincide at λ_c. Ierley and Ruehr also give asymptotic expansions for ϕ, which agree well with numerically computed solutions. The graphs of the expansion solutions are also in [11]. A subsequent paper by Troy [13] presents an analytic proof of the existence of at least one solution of the equations in the special case $\lambda > (27/4)^{1/3}$, for which the solutions must be asymptotically monotone.

The analysis in [11] for the boundary value problem (1)–(3) is analogous, using $\phi'(0) = \beta$. For $\lambda > 0$ there are two curves parametrically representing solutions; one with $\beta > 0$, the other with $\beta < 0$; see Fig. 4 in [11]. In [13], Troy proved analytically the existence of least one solution with $\beta = \phi'(0) > 0$ in the special case $\lambda > (27/4)^{1/3}$.

Section 2 of the present paper rigorously establishes the existence of a solution to (1)–(2) for *all* $\lambda > 0$ by a single geometric or dynamical systems method. Section 3 of the paper establishes the existence of at least one solution to (1)–(3) for all $\lambda > 0$. The proofs also give some bounds on the solutions as a secondary consequence of the geometric techniques.

The techniques of dynamical systems are useful for the analysis of nonlinear boundary value problems. A number of researchers have used these geometric techniques for a variety of nonlinear problems. For representative examples, see [3], [9], [10], [13], and [14]. Furthermore, the techniques employed here may give some insights into similar problems, e.g., the Falkner–Skann equation; see [8].

Dynamical systems methods translate differential equation problems into equivalent phase space problems in order to use the structure of the vector field, and the properties of solution curves. In particular, arguments in three-dimensional phase space are geometrically clear and easily motivated. Additional structure in the problem, such as invariant manifolds, is also revealed. The global structure of the invariant manifold can show where solutions to various boundary value problems occur. The variation of the manifold as problem parameters vary shows how solutions to boundary value problems can appear, disappear, coalesce, and vary with parameters. This is the approach adopted here. All arguments here could be expressed purely analytically, but they would lose motivation and geometric clarity. Moreover, no ingenious changes of independent or dependent variable need to be made in order to reveal aspects of the solution.

An overview of the paper will more fully describe the dynamical systems approach. The basic idea is related to the shooting method, or more formally to Wazewski's theorem. To make the paper self-contained, §2 contains a useful version of Wazewski's theorem. The third-order differential equation (1) is converted to a system of three first-order nonlinear equations by setting $u(s) = \phi(s)$, $v(s) = \phi'(s)$, $w(s) = \phi''(s)$, giving

$$\begin{aligned} u' &= v, \\ v' &= w, \\ w' &= \lambda v^2 - \lambda u w + u - 1. \end{aligned} \tag{4}$$

The system has only one equilibrium point $(1,0,0)$. The linearization of the system about this point always has one positive eigenvalue, and (4) has a corresponding one-dimensional unstable manifold. There are always two eigenvalues with negative real part, and so there is always a corresponding two-dimensional stable manifold. The two-dimensional stable manifold acts as a separatrix between the opposite branches of the unstable manifold. The unstable manifold as a set is attracting because it is transverse to the stable manifold at the equilibrium point. In this system, solutions along the unstable manifold are unbounded. Heuristically, any bounded solution of the equation, in particular a solution which satisfies the boundary condition $\phi(\infty) = 1$, cannot approach the unstable manifold except at the equilibrium point $(1, 0, 0)$. See Fig. 1.

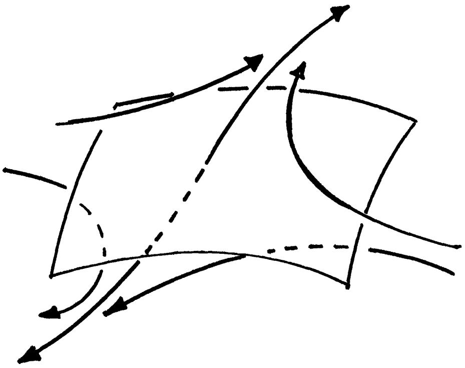


FIG. 1. *The orbits near the stable and unstable manifolds.*

An even simpler linear two-dimensional system analogous to (4) can convey the essential geometric elements of the argument. The analogous system is

$$\begin{aligned} u' &= 1 - u, \\ v' &= v, \end{aligned}$$

with boundary conditions $u(0) = 0$ and $u(\infty) = 1$. The equilibrium point is $(1, 0)$. The solutions of the system are $(1 + (u_0 - 1)e^{-t}, v_0 e^t)$. The unstable manifold is the line $u = 1$, and solutions starting at $(1, v_0)$ on this line will be unbounded. The stable manifold is the u-axis and all solutions starting at $(u_0, 0)$ will approach $(1, 0)$ as $t \to \infty$. Solutions starting at $(0, v_0)$ on the v-axis are $(1 - e^{-t}, v_0 e^t)$. All solutions (except when $v_0 = 0$) become unbounded as they approach the unstable manifold. The exception to becoming unbounded can be identified as the intersection of the stable manifold, along which solutions approach $(1, 0)$, with the initial condition set with coordinates $(0, v_0)$. See Fig. 2.

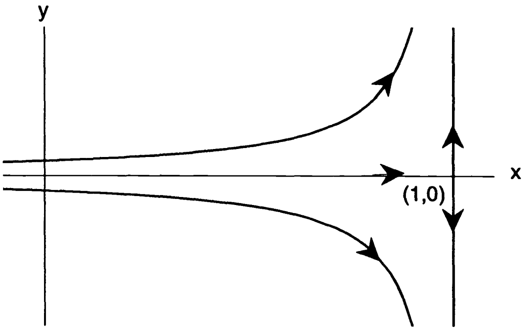


FIG. 2. *The simple two-dimensional analog system.*

This solution satisfies the boundary conditions and so is the solution we seek.

The situation for (4) is very similar but is complicated by the nonlinearity and the three-dimensional setting. The method is to use regions of space to capture the essence of the unboundedness along the unstable manifold. Removing a pair of regions in three-dimensional phase space suggested by the unboundedness of orbits approaching the unstable manifold defines a Wazewski set W which heuristically contains the bounded orbits. The immediate exit set of the Wazewski set W is disconnected. An interval Σ of initial values is defined. Orbits of the system starting at one end of the interval Σ will exit W into one of the removed regions, while orbits starting at the other end of Σ will exit W into the other region. By the connectedness of Σ, there must be an orbit starting on the interval that does not exit W. We show this orbit must remain bounded. Then, using the Lyapunov function

$$V_0(u,v,w) = \left(\frac{w^2}{2}\right) - \left(\frac{\lambda v^3}{3}\right) - (u-1)v,$$

with Lyapunov derivative

$$\dot{V}_0(u,v,w) = -\lambda u w^2 - v^2,$$

and the LaSalle Invariance Principle, this orbit must approach the equilibrium point $(1,0,0)$.

The LaSalle Invariance Principle weakens the requirements demanded of a Lyapunov function while still retaining a useful conclusion. This simplifies the task of finding a suitable Lyapunov function. The Lyapunov function $V_0(u,v,w)$ is not motivated by any energy function or physical reasoning. The Lyapunov function $V_0(u,v,w)$ and the related family of functions

$$V_\alpha(u,v,w) = \frac{w^2}{2} - \frac{\lambda v^3}{3} - (u-1)v + \alpha\left(vw + \frac{\lambda u v^2}{2} - \frac{(u-1)^2}{2}\right),$$

with Lyapunov derivative

$$\dot{V}_\alpha(u,v,w) = (\alpha - \lambda u)w^2 + \left(\frac{3\alpha\lambda v}{2} - 1\right)v^2,$$

play an important role in the paper. The family of Lyapunov functions V_α suggest yet another useful function

$$V_\infty(u,v,w) = vw + \frac{\lambda u v^2}{2} - \frac{(u-1)^2}{2},$$

with Lyapunov derivative

$$\dot{V}_\infty(u,v,w) = w^2 + \frac{3\lambda v^3}{2}.$$

It is worth describing how such a family of Lyapunov functions can be obtained. The equation system has simple polynomial terms so that "trial and error" arrangement and cancellation of the terms easily yields a Lyapunov derivative

$$\dot{V}_0(u,v,w) = -\lambda u w^2 - v^2,$$

which is negative definite for $u \geq 0$. Likewise, more "trial and error" arrangement of the derivative terms of the system leads to the Lyapunov derivative

$$\dot{V}_\alpha(u,v,w) = (\alpha - \lambda u)w^2 + \left(\frac{3\alpha\lambda v}{2} - 1\right)v^2,$$

which will be negative definite along bounded orbits with $u \geq 0$ for $\alpha < 0$ sufficiently small.

Alternatively, multiply (1) by ϕ'' and integrate from zero to s to obtain

$$\frac{(\phi''(s))^2}{2} - \frac{(\phi''(0))^2}{2} + \int_0^s \lambda\phi(z)(\phi''(z))^2\,dz - \frac{\lambda(\phi'(s))^3}{3} + \frac{\lambda(\phi'(0))^3}{3}$$
$$+\phi'(s) - \phi'(0) - \int_0^s \phi(z)\phi''(z)\,dz = 0.$$

Integrate the second integral by parts, rearrange, and use $\phi(0) = 0$, $\phi'(0) = 0$, $\phi''(0) = \beta$, to obtain

$$\frac{(\phi''(s))^2}{2} - \frac{\beta^2}{2} - \frac{\lambda(\phi'(s))^3}{3} + \phi'(s) - \phi(s)\phi'(s) = -\int_0^s \lambda\phi(z)(\phi''(z))^2\,dz - \int_0^s (\phi'(z))^2\,dz.$$

The right side of the equation is clearly decreasing on intervals where $\phi(s) > 0$. Converting the left side of the equation to u, v, w suggests the Lyapunov function V_0.

A third approach is to use the general procedure for constructing Lyapunov equations for third-order equations in [1]. For the general Lyapunov function given in [1] choosing $\alpha = 0$ and the weight function to be identically 1 yields the function V_0. Using the constant $\alpha \neq 0$ and the weight function to be identically 1 supplies V_α.

2. Existence of solutions to the no-slip problem. This section contains the proof of the following.

THEOREM 1. *For all $\lambda > 0$ there is at least one solution of the problem*

$$\phi''' + \lambda(\phi\phi'' - (\phi')^2) + 1 - \phi = 0, \tag{5}$$

with the "no-slip" boundary conditions

$$\begin{aligned} \phi(0) &= 0, \\ \phi'(0) &= 0, \\ \phi(\infty) &= 1. \end{aligned} \tag{6}$$

We begin the proof by converting (5) to a system of 3 first-order equations by setting $u = \phi$, $v = \phi'$, and $w = \phi''$. Here $' = d/ds$, so that s is the independent variable. Occasionally, in the spirit of dynamical systems theory, we will refer to the independent variable s as "time" even though the independent variable actually represents a spatial extent in the similarity solution. The equation (5) is equivalent to the system

$$\begin{aligned} u' &= v, \\ v' &= w, \\ w' &= \lambda v^2 - \lambda uw + u - 1. \end{aligned} \tag{7}$$

The following proposition is a variant of Wazewski's theorem, which is a formalization and extension of the "shooting method." This proposition recognizes that the flow defined by the solutions of a differential equation gives a topological mapping between regions of phase space. The statement is the same as in [4], where the proof is also given. Both the statement and the proof are variants of those given in [2]. The notation is the same as that in [2].

Consider the autonomous differential equation

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}), \qquad \mathbf{y} \in \mathbf{R}^n, \quad ' = \frac{d}{ds}, \tag{8}$$

where $\mathbf{f} : \mathbf{R}^n \to \mathbf{R}^n$ is continuous and satisfies a Lipschitz condition. Let $\mathbf{y}(s; \mathbf{y}_0)$ be the unique solution of (8) satisfying $\mathbf{y}(0; \mathbf{y}_0) = \mathbf{y}_0$. For convenience set $\mathbf{y}(s; \mathbf{y}_0) = \mathbf{y}_0 \cdot s$. Let $U \cdot S$ be the set of points $\mathbf{y}_0 \cdot s$, where $\mathbf{y}_0 \in U$, $s \in S$.

In order to state the proposition some definitions are necessary. Given $W \subset \mathbf{R}^n$, set

$$W^- = \{\mathbf{y}_0 \in W : \text{any } s > 0, \quad \mathbf{y}_0 \cdot [0, s) \not\subset W\}.$$

W^- is called the *immediate exit set* of W. The continuity of $\mathbf{y}_0 \cdot s$ in s implies that interior points of W are not in the immediate exit set, or equivalently $W^- \subset (\partial W \cap W)$. Given $\Sigma \subset W$, let

$$\Sigma^0 = \{\mathbf{y}_0 \in \Sigma : \text{ there is an } s_0 \text{ such that } \mathbf{y}_0 \cdot s_0 \notin W\}.$$

For $\mathbf{y}_0 \in \Sigma^0$ define

$$T(\mathbf{y}_0) = \sup\{s : \mathbf{y}_0 \cdot [0, s] \subset W\}.$$

$T(\mathbf{y}_0)$ is called an *exit time*. Note that if $\mathbf{y}_0 \cdot T(\mathbf{y}_0) \in W$, then $\mathbf{y}_0 \cdot T(\mathbf{y}_0) \in W^-$. Note also that for $\mathbf{y}_0 \in W$, $T(\mathbf{y}_0) = 0$ if and only if $\mathbf{y}_0 \in W^-$. The notation $\mathrm{cl}(W)$ is used for the closure of W.

PROPOSITION 2. *Suppose*

(1) *If* $\mathbf{y}_0 \in \Sigma$ *and* $\mathbf{y}_0 \cdot [0, s_0] \subset \mathrm{cl}(W)$, *then* $\mathbf{y}_0 \cdot [0, s_0] \subset W$;

(2) *If* $\mathbf{y}_0 \in \Sigma$, $\mathbf{y}_0 \cdot [0, s_0] \subset W$, *and* $\mathbf{y}_0 \cdot s \notin W^-$ *for each* $s \in [0, s_0]$, *then there is an open set* V_s *about* $\mathbf{y}_0 \cdot s$ *disjoint from* W^-;

(3) $\Sigma = \Sigma^0$, Σ *is compact, and* Σ *intersects an orbit of* (8) *only once.*

Then the mapping $F(\mathbf{y}_0) = \mathbf{y}_0 \cdot T(\mathbf{y}_0)$ *is a homeomorphism from* Σ *to its image on* W^-.

The proof is in [4]. Notice that the first hypothesis is trivially satisfied if the set W is closed. The second hypothesis is satisfied if orbits from Σ have no tangencies to the boundary of W within W, i.e., no *internal tangencies.* The set Σ corresponds to an interval of initial conditions, and the third hypothesis is a technical hypothesis which will be satisfied in all reasonable cases.

The system (7) has just one equilibrium point, namely (1,0,0). It is easy to check that the Jacobian matrix of the linearization at (1,0,0) is

$$J(1,0,0) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -\lambda \end{pmatrix}.$$

The characteristic equation is

$$-r^3 - \lambda r^2 + 1 = 0.$$

Using the Routh–Hurwitz criterion (see [5]), it is easy to show that for all values of λ there is always one positive eigenvalue r_1 and two eigenvalues with negative real part. If $\lambda \geq (27/4)^{1/3}$, there are three real roots. If $\lambda < (27/4)^{1/3}$, there is a complex conjugate pair of eigenvalues with negative real part. Label the roots so that $r_1 > 0 > \mathrm{Re}(r_2) \geq \mathrm{Re}(r_3)$.

Consider the eigenvector $(1, r_1, r_1^2)^T$ (where T is used for the transpose) corresponding to the positive eigenvalue r_1. The unstable manifold of the equilibrium point $(1,0,0)$ is tangent to the eigenvector corresponding to the positive eigenvalue [7]. In fact, due to the special nature of the nonlinearity in (7), the solutions on the unstable manifold are precisely $(1,0,0)^T + C(1, r_1, r_1^2)^T e^{r_1 s}$, where C is an arbitrary nonzero constant; so the unstable manifold is the straight line through $(1,0,0)$ in the direction $(1, r_1, r_1^2)^T$. Therefore, the octant

$$\mathcal{O}_1 = \{(u,v,w) : u > 1, v > 0, w > 0\}$$

contains one branch of the unstable manifold. The other branch of the unstable manifold enters the region $\mathcal{O}_2 = \{(u,v,w) : u < 0\}$. The regions $\mathcal{O}_1$ and $\mathcal{O}_2$ surround most of the unstable manifold. Technical reasons related to the second hypothesis in Proposition 2 and the necessity to have the Lyapunov derivative of V_0 negative definite compel the choice $\mathcal{O}_2 = \{(u,v,w) : u < 0\}$ rather than the more obvious choice $\{(u,v,w) : u < 1, v < 0, w < 0\}$, which completely contains the branch of the unstable manifold in $\{u < 1\}$.

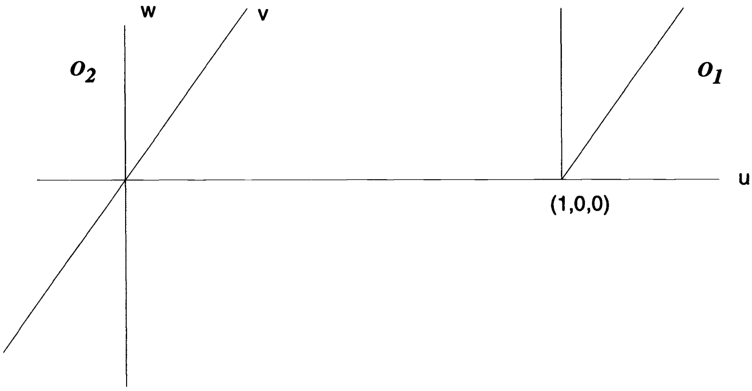


FIG. 3. *The regions $\mathcal{O}_1$ and $\mathcal{O}_2$.*

Define the Wazewski set $W = R^3 - (\mathcal{O}_1 \cup \mathcal{O}_2)$; see Fig. 3. The closed set W is the exterior of the disjoint regions and contains the local stable manifold of the point $(1,0,0)$. Since W is closed, hypothesis 1 of Proposition 2 is satisfied. Although the region W is the Wazewski set, it is more convenient and natural to work with the removed regions $\mathcal{O}_1$ and $\mathcal{O}_2$. Thus, much of the following will be done in terms of $\mathcal{O}_1$ and $\mathcal{O}_2$ and then referred back to the main object of interest, which is W.

In order to apply Proposition 2, we need to examine the immediate exit set W^-. This will show that the exit set is disconnected. First examine the vector field on the boundary of $\mathcal{O}_1$. This determines where an orbit that hits this portion of the boundary of W will go.

(1) Verification that the side, $v = 0$, is an exit set portion: On the plane quadrant $u > 1, v = 0, w > 0$, we have $v' = w > 0$ so that an orbit hitting this portion of the

boundary of W will immediately enter the octant $\mathcal{O}_1$. Therefore, the plane quadrant $\{u > 1, v = 0, w > 0\} \subset W^-$. We abbreviate this by saying the set is an exit set portion.

(2) Verification that the front, $u = 1$, is an exit set portion: On the plane quadrant $u = 1$, $v > 0$, $w > 0$, we have $u' = v > 0$.

(3) Verification that the bottom, $w = 0$, is an exit set portion: On the plane quadrant $u > 1$, $v > 0$, $w = 0$, we have $w' = \lambda v^2 - \lambda uw + u - 1 > 0$.

(4) Verification that the edge $u = 1$, $v = 0$, $w > 0$ is an exit set portion: On this edge, $u' = v = 0$, and $u'' = v' = w > 0$ so that an orbit touching this edge has a minimum in the u coordinate and has the v coordinate increasing. The orbit exits W into the octant $\mathcal{O}_1$.

(5) Verification that the edge $u > 1$, $v = 0$, $w = 0$ is an exit set portion: On this edge, $v' = w = 0$, $v'' = w' = \lambda v^2 - \lambda uw + u - 1 > 0$, so that v has a local minimum here, and $w' = u - 1 > 0$, and of course $u > 1$, so that an orbit touching this edge immediately enters the octant $\mathcal{O}_1$.

(6) Verification that the edge $w = 0$, $u = 1$, $v > 0$ is an exit set portion: On this edge, $w' = \lambda v^2 - \lambda uw + u - 1 > 0$ and $u' = v > 0$. Therefore any orbit touching this edge must enter the octant $\mathcal{O}_1$.

Next, examine the vector field on the boundary of $\mathcal{O}_2$.

(1) Verification that the face $u = 0$, $v < 0$ is an exit set portion: Here we have $u' = v < 0$ and so orbits touching this half of the face of $\mathcal{O}_2$ must immediately exit into $\mathcal{O}_2$.

(2) The face $u = 0$, $v > 0$ is *not* an exit set portion: Here we have $u' = v > 0$ and so orbits are actually entering the set W from $\mathcal{O}_2$.

(3) Verification that the negative w-axis, $u = 0$, $v = 0$, $w < 0$ is an immediate exit set portion: On this portion of the face of $\mathcal{O}_2$ an orbit has $u' = v = 0$, $u'' = v' = w < 0$ so that the u coordinate has a local maximum, and the orbit re-enters the region $\{u < 0\}$.

(4) The positive w-axis, $u = 0$, $v = 0$, $w > 0$ is *not* an immediate exit set portion: On this portion of the face of $\mathcal{O}_2$ an orbit has $u' = v = 0$, $u'' = v' = w > 0$ so that the u coordinate has a local minimum and the v-coordinate is increasing. Orbits passing through points on the positive w-axis remain in the Wazewski set W.

(5) The origin $(0, 0, 0)$ is part of the exit set since an orbit from this point has $u' = v = 0$, $u'' = v' = w = 0$, and $u''' = w' = -1$. Thus the orbit has a cubic singularity in u at the origin, and immediately enters $\mathcal{O}_2$.

Therefore, all of the boundary of the octant $\mathcal{O}_1$ except for the corner equilibrium point $(1, 0, 0)$ is in the exit set W^-. The portion of the exit set W^- on the boundary of $\mathcal{O}_2$ is an open half-plane $\{u = 0, v < 0\}$, together with the nonpositive w-axis. See Fig. 4. In fact, orbits in W may have internal tangencies to the boundary of W along the positive w-axis. Lemma 5 will show that orbits starting from the interval of initial conditions Σ cannot have these internal tangencies on the positive w-axis so hypothesis 2 of Proposition 2 will still be satisfied. Referring back to the Wazewski set W, the exit set W^- is obviously a disconnected set.

Next we define the set Σ needed by Proposition 2 to prove Theorem 1. The no-slip initial conditions set Σ will be a compact interval on the w-axis. Defining Σ requires some technical lemmas.

Lemma 3 shows solutions with $u(0) = 0$ and $v(0) = 0$ cannot grow unboundedly in the positive v or positive w directions without entering the octant $\mathcal{O}_1$.

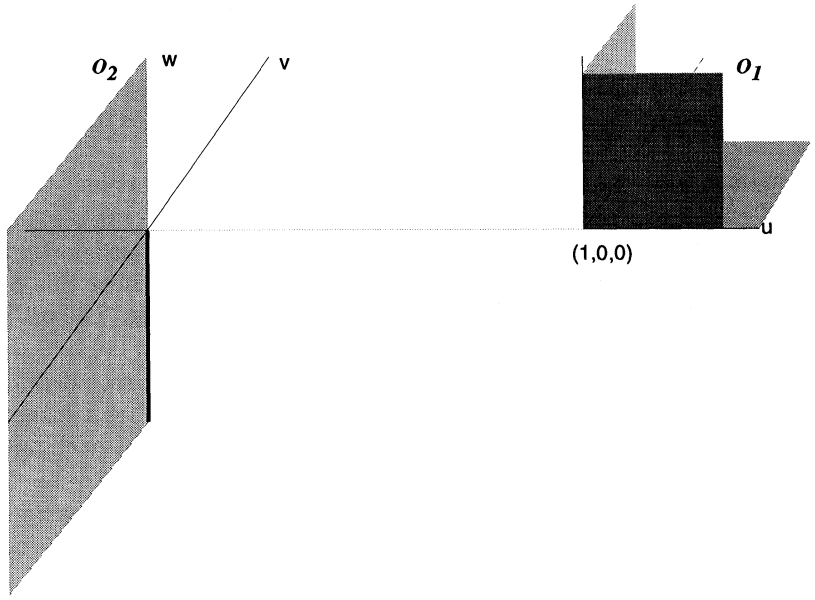


FIG. 4. *The exit set W^-.*

Some notation is necessary. Fix $w_0 > 0$, and let

$$u_0(v) = \int_0^v \frac{z}{\sqrt{2\lambda z^3/3 + w_0^2}}\, dz.$$

The integral clearly exists for all $v \geq 0$ and $u_0(v)$ is increasing. According to Gradshteyn and Ryzhik [6, §2.202, p. 71] it is not possible to express $u_0(v)$ in terms of the elementary functions. Let $v_0(u)$ be the inverse function. Additionally, let $\Lambda = \max(\lambda, 1/\sqrt{\lambda})$. Define the box-like region

$$R = \{(u, v, w) : 0 \leq u \leq 1, 0 \leq v \leq v_0(u) + \Lambda, 0 \leq w \leq (v_0(1) + \Lambda)v + w_0\};$$

see Fig. 5.

LEMMA 3. *An orbit of* (7) *starting from the point* $(0, 0, w(0))$ *with* $0 \leq w(0) \leq w_0$ *cannot leave the bounded region* R *except transversally through the floor* $0 \leq u \leq 1$, $0 \leq v \leq v_0(u) + \Lambda$, $w = 0$, *or transversally through the back wall* $u = 1$, $0 \leq v \leq v_0(1) + \Lambda$, $0 \leq w \leq (v_0(1) + \Lambda)v + w_0$ *into the octant* $\mathcal{O}_1$.

Proof. The proof proceeds by examining the vector field on the sides of the box-like region. The vector field points inward on some of these sides, so that an orbit cannot leave through these sides.

(Front Wall.) On the edge $u = 0$, $v = 0$, $0 < w \leq w_0$, orbits have $u' = 0$ but $u'' = v' > 0$. Therefore orbits starting on the edge actually enter R with a quadratic tangency. In the region R, the orbits have $u' = v > 0$; so the orbits cannot reverse direction and exit through the front wall $u = 0$, $0 \leq v \leq \Lambda$, $0 \leq w \leq (v_0(1) + \Lambda)v + w_0$.

(Side wall.) In the region R the orbits have $v' = w > 0$; so the orbits are increasing in v from zero and orbits cannot leave through the wall $0 \leq u \leq 1$, $v = 0$, and $0 \leq w \leq w_0$.

(Curved side wall.) The orbit has $w' = \lambda v^2 - \lambda uw + u - 1$, so that as long as the orbit is in the region where $0 \leq u < 1$, and $w > 0$, we know that $w' < \lambda v^2$. Now

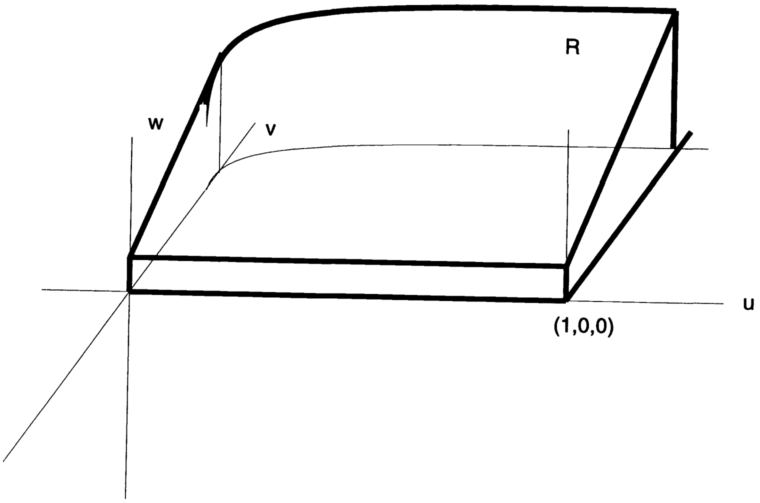


FIG. 5. *The bounded region R.*

$w' = v''$, so that this can be written as $v'' < \lambda v^2$. Multiply through by $v' = w > 0$ and integrate from zero to s to obtain

$$\frac{(v')^2}{2} - \frac{w(0)^2}{2} < \frac{\lambda v^3}{3}.$$

Rearrange and use $w(0) \leq w_0$ to get

$$v' < \sqrt{2\lambda v^3/3 + w_0^2}.$$

Therefore, orbits of the system (7) must have

$$u' = v,$$
$$v' < \sqrt{2\lambda v^3/3 + w_0^2}.$$

The outward normal (in the direction of increasing v) of the curved side wall $v = v_0(u) + \Lambda$, $0 \leq u < 1$, $0 < w \leq (v_0(1) + \Lambda)v + w_0$ is $(-\sqrt{2\lambda v^3/3 + w_0^2}/v, 1, 0)$. The dot product of the tangents to the orbits and the outward normal of the surface is therefore negative, and orbits cannot cross this side wall.

Because $\Lambda \geq 1/\sqrt{\lambda}$ and because $v_0(u)$ is increasing, the bottom edge of the curved side wall, $v = v_0(u) + \Lambda$, $0 < u \leq 1$, $w = 0$ lies in the region where $w' = \lambda v^2 - \lambda uw + u - 1 > 0$. Therefore, orbits from the initial condition set cannot leave the region R from the bottom edge of the curved side wall either.

(Ceiling.) For the orbits starting from $(0, 0, w(0))$ with $0 \leq w(0) \leq w_0$, we have $v'(0) = w(0) \geq 0$ and $w'(0) = -1 \leq 0$. Thus the orbit starts below the upward sloping open ceiling panel $w = (v_0(1) + \Lambda)v + w_0$, $0 < u < 1$, $0 \leq v \leq (v_0(1) + \Lambda)$. Suppose there is a first time s when the orbit hits the upward sloping ceiling panel:

$$w(s) = (v_0(1) + \Lambda)v(s) + w_0.$$

Then

$$w'(s) \geq (v_0(1) + \Lambda)v'(s),$$

or equivalently,

$$\lambda v^2 - \lambda uw + u - 1 \geq (v_0(1) + \Lambda)w.$$

Now substituting $w = (v_0(1) + \Lambda)v + w_0$ on the right:

$$\lambda v^2 - \lambda uw + u - 1 \geq (v_0(1) + \Lambda)^2 v + (v_0(1) + \Lambda)w_0. \tag{9}$$

But $\lambda v^2 \leq (v_0(1) + \Lambda)^2 v$ on the region $0 \leq v \leq v_0(1) + \Lambda$ and the remaining terms on the left-hand side of (9) are nonpositive and the remaining term on the right side of (9) is strictly positive which gives a contradiction. Alternatively, this argument says that the projection of the vector field on the normal vector of the ceiling plane is negative. That is, the vector field on the ceiling of the box points into the box and the orbit cannot leave the box through the ceiling. □

LEMMA 4. *Let*

$$A(\lambda) = \begin{cases} \sqrt{2}\lambda^{-1/4} & \text{for } 0 < \lambda < 2^{2/3}, \\ \sqrt{\lambda} & \text{for } 2^{2/3} \leq \lambda. \end{cases}$$

An orbit from $(0, 0, w(0))$ *with* $w(0) > A(\lambda)$ *cannot leave the bounded region* R *through the floor* $w = 0$. *Consequently, orbits from the* w*-axis with* $w(0) > A(\lambda)$ *can only leave* R *through the back wall* $u = 1$ *into the octant* $\mathcal{O}_1$.

Proof. Observe that $w' > 0$ when $u > 1 - \lambda v^2$ and $w = 0$. Thus, an orbit cannot leave the region R given in Lemma 3 through the floor outside the parabolic half-cylinder $u = 1 - \lambda v^2$ with $v \geq 0$.

Note that u and v are increasing for an orbit in R. Therefore, it remains to show that an orbit from $(0, 0, w(0))$ with $w(0) > A(\lambda)$ cannot go through the floor within the parabolic half-cylinder. In fact we will show that all the more so, the orbit cannot cross the downward sloping plane panel $w = A(\lambda) - \sqrt{\lambda}A(\lambda)v$ within the parabolic half-cylinder. See Fig. 6 for a diagram of the sloping floor panel.

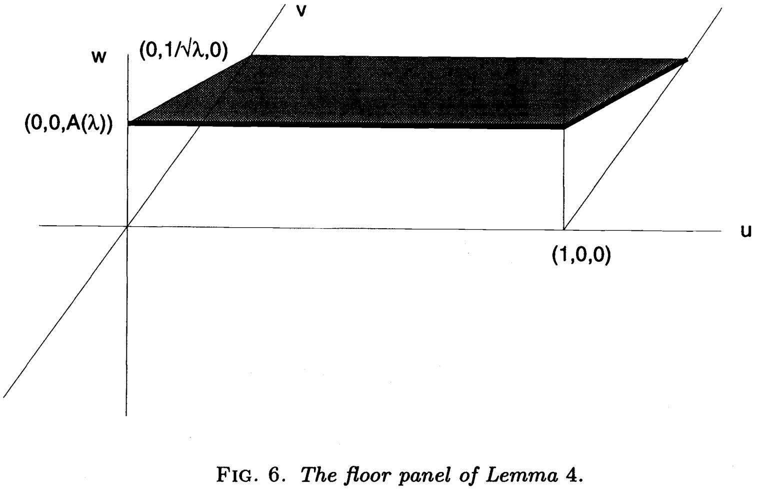


FIG. 6. *The floor panel of Lemma* 4.

At the initial time $s = 0$,

$$w(0) > A(\lambda) - \sqrt{\lambda}A(\lambda)v(0).$$

Suppose there is a first time s the orbit touches the downward sloping floor panel so that

$$w(s) = A(\lambda) - \sqrt{\lambda}A(\lambda)v(s).$$

Then $w'(s) + \sqrt{\lambda}A(\lambda)v'(s) \leq 0$. If this inequality is impossible within the parabolic half-cylinder, then the orbits cannot touch this plane panel.

Substitute the expressions from the differential equations and the equation of the plane into $w' + \sqrt{\lambda}A(\lambda)v'$ to obtain

$$-1 + \sqrt{\lambda}A^2 + (1 - \lambda A)u - \lambda A^2 v + (\lambda)^{3/2}Auv + \lambda v^2. \tag{10}$$

If the minimum value of (10) over the half parabola $0 \leq u \leq 1 - \lambda v^2$, $0 \leq v$ is positive, the proof is done. The expression (10) has only one critical point, which is easily seen to be a saddle point. Thus, if the values of (10) on the boundary of the half parabola over which the panel is defined are nonnegative, an orbit cannot cross the plane panel in the interior of the half parabola.

Consider first the case when $\lambda \geq 2^{2/3}$, so $A = \sqrt{\lambda}$. Over the edge $0 \leq u \leq 1$, $v = 0$ the value of (10) is

$$(\lambda^{3/2} - 1)(1 - u),$$

which is clearly nonnegative.

The value of (10) over the edge $u = 0$, $0 \leq v \leq 1/\sqrt{\lambda}$ is

$$-1 + \lambda v^2 - \lambda^2 v + \lambda^{3/2},$$

which is also nonnegative.

The value of (10) along the parabolic edge $u = 1 - \lambda v^2$, $0 \leq v \leq 1/\sqrt{\lambda}$ is

$$\lambda^{5/2}v^2(1 - \sqrt{\lambda}v),$$

also clearly nonnegative.

Now for the case $0 < \lambda < 2^{2/3}$, the expression for the projection is

$$1 + (1 - \sqrt{2}\lambda^{3/4})u - 2\sqrt{\lambda}v + \sqrt{2}\lambda^{5/4}uv + \lambda v^2. \tag{11}$$

Over the edge $0 \leq u \leq 1$, $v = 0$ the value of (11) is

$$1 + (1 - \sqrt{2}\lambda^{3/4})u,$$

which is clearly nonnegative. The value of (11) over the edge $u = 0$, $0 \leq v \leq 1/\sqrt{\lambda}$ is

$$(1 - \sqrt{\lambda}v)^2,$$

which is also nonnegative. Finally, the value of (11) along the parabolic edge $u = 1 - \lambda v^2$, $0 \leq v \leq 1/\sqrt{\lambda}$ is

$$(2 - \sqrt{2}\lambda^{3/4} + \sqrt{2}\lambda^{7/4}v^2)(1 - \sqrt{\lambda}v),$$

also clearly nonnegative. □

Remark. In Lemma 1 of [13], Troy obtained the uniform (in $\lambda > 0$) estimate that all orbits starting with $w(0) > 7/3$ enter $\mathcal{O}_1$. Lemma 4 provides an improvement on this estimate for the range $4(3/7)^4 < \lambda < (7/3)^2$. This range covers much of the interval for which the results are new.

LEMMA 5. *Define*

$$\Sigma = \{(0,0,\beta) : 0 \le \beta \le A(\lambda) + 1\},$$

where $A(\lambda)$ is the constant defined by Lemma 4. Then orbits starting from the initial condition set Σ cannot have internal tangencies in the Wazewski set W.

Proof. The proof will be accomplished by examining several cases. The case-by-case examination and variety of arguments among the cases seems to be required because the proof necessarily determines the long-time nonmonotone behavior of a family of solutions to a nonlinear initial value problem.

From the previous determination of the exit set W^-, the only possible place for an orbit to have an internal tangency in W is on the positive w-axis on the upper face of the region $\mathcal{O}_2$. Therefore, the proof will follow points from Σ forward in time and show that the orbits cannot touch the positive w-axis.

Some initial notation is necessary to establish the cases. Suppose, for a contradiction, that there is an orbit from Σ with an internal tangency. That is, suppose that there is an orbit from $(0, 0, w(0))$ that enters the region $u > 0$, crosses the u-w plane at $(u(T_2), 0, w(T_2))$ for some $T_2 > 0$ and then decreases to $(0, 0, w(T_7))$ for some $T_7 > T_2$, where $w(T_7) > 0$. In Fig. 7 is a diagram of the u-v projection of such a hypothesized orbit, together with the labeling of the important points. For the rest of the proof fix attention on this orbit. The cases will depend on the value of $u(T_2)$. (The notation for the times T_2, T_7 allows for additional intermediate times to be defined later.)

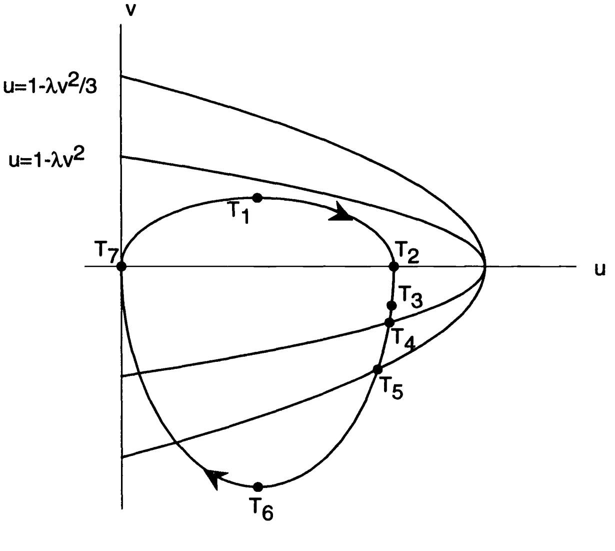


FIG. 7. *Diagram of the u-v projection of the orbit of the proof of Lemma 5.*

It will be convenient to use the abuse of notation $V_0(s) = V_0(u(s), v(s), w(s))$.

Case 1. $u(T_2) < 1$

Subcase (a). $\lambda \le (27/16)^{2/3} \approx 1.4174$.

The orbit from $(0, 0, w(0))$ enters the region $u > 0$ and crosses the floor $w = 0$ at T_1 in the region $0 < u < 1$, $u < 1 - \lambda v^2$. The orbit then decreases in v, until crossing the u-w plane with $u(T_2) < 1$ and $w(T_2) < 0$. The orbit continues decreasing in v until recrossing the u-v plane at $(u(T_6), v(T_6), 0)$ where $v(T_6) < 0$, and then the orbit approaches the point of internal tangency $(0, 0, w(T_7))$.

From the Lyapunov function $V_0(u, v, w) = w^2/2 - \lambda v^3/3 - (u-1)v$ we obtain that $V_0(s) > V_0(T_7) = (w(T_7))^2/2 > 0$ for $s < T_7$. In particular, $V_0(T_6) = -\lambda v(T_6)^3/3 - (u(T_6) - 1)v(T_6) > 0$ so that $u(T_6) > 1 - \lambda v(T_6)^2/3$. This means the orbit must cross the parabolic cylinder $u = 1 - \lambda v^2/3$ at some T_5. Therefore, at T_5, we have $u' + 2\lambda v v'/3 = v + 2\lambda v w/3 \geq 0$. So at T_5, $w \leq -3/(2\lambda)$. At T_5, $V_0(T_5) = w(T_5)^2/2 \geq 9/(8\lambda^2)$. Therefore, at T_2, $V_0(T_2) \geq w(T_2)^2/2 > 9/(8\lambda^2)$, and so we conclude that $w(T_2) < -3/(2\lambda)$.

Now at T_1 the orbit is in the region $w = 0$, $u > 0$, $v > 0$, and $u < 1 - \lambda v^2$. The maximum value of $V_0(u, v, w)$ on this region is $2/(3\sqrt{\lambda})$ occurring at $(1/\sqrt{\lambda}, 0, 0)$. Therefore $V_0(T_2) < V_0(T_1) \leq 2/(3\sqrt{\lambda})$. But for $\lambda \leq (27/16)^{2/3}$, $9/(8\lambda^2) \geq 2/(3\sqrt{\lambda})$; so we have a contradiction and no internal tangency is possible.

Subcase (b). $\lambda > (27/16)^{2/3}$.

For this case, we need some additional bounds on the orbit. These bounds will show that it is not possible for the orbit to loop around the "bubble" defined by $V_0(u, v, w) = 0$ in the region $v < 0$.

The previous argument shows that on the interval $0 < s < T_2$, where $v > 0$ an orbit with an internal tangency, must decrease in w to $w(T_2) < -3/(2\lambda)$. From the vector field this can only happen if the orbit passes through the half parabolic region $w = -3/(2\lambda)$, $v > 0$ and $0 < u < (2/5)(1 - \lambda v^2)$. By examining the vector field on the isocline surface $w' = \lambda v^2 - \lambda u w + u - 1 = 0$, we see that the orbit can pass from the region $w' < 0$ to $w' > 0$ in the half space $v > 0$ only where $w \geq -1/\lambda$. Therefore, at $s = T_2$ the orbit must still be in the region $w' < 0$ so $-\lambda u(T_2) w(T_2) + u(T_2) - 1 < 0$. Since $w(T_2) < -3/(2\lambda)$ we must have that $u(T_2) < 2/5$. Thereafter, in the region $v < 0$, certainly $u < 2/5$.

The previous subcase showed that $V_0(s) > 0$ for all $s \leq T_7$. Let T_4 be the time when the orbit crosses the parabolic cylinder $u = 1 - \lambda v^2$ with $v < 0$. Then $V_0(T_4) = w^2(T_4)/2 + 2\lambda v^3(T_4)/3 > 0$ or $w(T_4) < -\sqrt{4\lambda/3}(-v)^{3/2}$. In terms of u along $u = 1 - \lambda v^2$, we have $w < -\sqrt{4/3}(1-u)^{3/4}/\lambda^{1/4}$. The orbit has its minimum value of w at T_3 when the orbit crosses the $w' = 0$ isocline surface $\lambda v^2 - \lambda u w + u - 1 = 0$. At $s = T_3$ the u and w coordinates of the orbit satisfy $\lambda v^2 - \lambda u w + u - 1 = 0$, so $w > (1/\lambda)(1 - 1/u)$. For $s > T_3$ the w coordinate is increasing and the u coordinate is decreasing. In particular, $w(T_4) > (1/\lambda)(1 - 1/u(T_4))$ and we have all the more so that at T_4,

$$\frac{1}{\lambda}\left(1 - \frac{1}{u(T_4)}\right) < -\sqrt{\frac{4}{3}}\frac{(1 - u(T_4))^{3/4}}{\lambda^{1/4}}$$

Solving for u we obtain the estimate that $u(T_4) < (\sqrt{3}/2)\lambda^{-3/4}$.

Combining the arguments of the previous two paragraphs, we find that

$$u(T_4) < \min\left(\frac{2}{5}, \left(\frac{\sqrt{3}}{2}\right)\lambda^{-3/4}\right) \equiv u_4$$

and

$$\frac{1}{\lambda}\left(1 - \frac{1}{u_4}\right) < w(T_4) < -\sqrt{\frac{4}{3}}\frac{(1 - u_4)^{3/4}}{\lambda^{1/4}}. \tag{12}$$

Note that $u_4 = 2/5$ for $\lambda \leq (75/16)^{2/3} \approx 2.8009$.

Next we show that the orbit cannot cross the plane

$$w = -2\sqrt{\frac{4}{3}}\frac{(1-u_4)^{3/4}}{u_4\lambda^{1/4}}u + \sqrt{\frac{4}{3}}\frac{(1-u_4)^{3/4}}{\lambda^{1/4}}. \tag{13}$$

The previous estimates (12) on the u and w coordinates show that the orbit is below this plane at $s = T_4$. If for some $s > T_4$ the orbit crosses the plane, then at that time we would have that

$$w' + 2\sqrt{\frac{4}{3}}\frac{(1-u_4)^{3/4}}{u_4\lambda^{1/4}}u' \geq 0$$

or

$$\lambda v^2 - \lambda uw + u - 1 + 2\sqrt{\frac{4}{3}}\frac{(1-u_4)^{3/4}}{u_4\lambda^{1/4}}v \geq 0.$$

By factoring and rearranging, this can be written as

$$\sqrt{\lambda}v\left(\sqrt{\lambda}v + 2\sqrt{\frac{4}{3}}\frac{(1-u_4)^{3/4}}{u_4\lambda^{3/4}}\right) - \lambda uw + u - 1 \geq 0.$$

At this time, we know that $\sqrt{\lambda}v < 0$ and $-\lambda uw + u - 1 < 0$. Thus, if we show that

$$\sqrt{\lambda}v + 2\sqrt{\frac{4}{3}}\frac{(1-u_4)^{3/4}}{u_4\lambda^{3/4}} \geq 0,$$

we will have reached the desired contradiction showing that the orbit cannot cross the plane.

To do this, we need a bound on the v coordinate. Now v has its minimum negative value at $s = T_6$ when $w = 0$. As shown above, $V_0(T_6) = -\lambda v(T_6)^3/3 - (u(T_6) - 1)v(T_6) < 2/(3\sqrt{\lambda})$. Since $0 \leq u(T_6) < 1$, it is easy to see that $v(T_6) > -2^{1/3}/\sqrt{\lambda}$. Therefore, we obtain the bound $v(s) > -2^{1/3}/\sqrt{\lambda}$. Then the desired contradiction can be obtained by showing that

$$-2^{1/3} + 2\sqrt{\frac{4}{3}}\frac{(1-u_4)^{3/4}}{u_4\lambda^{3/4}} \geq 0$$

for the values of $\lambda > (27/16)^{2/3}$ under consideration. For $(27/16)^{2/3} < \lambda \leq (75/16)^{2/3}$, $u_4 = 2/5$ and the calculation reduces to showing that

$$\lambda^{3/4} \leq \left(\frac{3}{5}\right)^{3/4}\frac{10}{2^{1/3}\cdot 3^{1/2}},$$

which is true for this range of λ.

For $\lambda > (75/16)^{2/3}$, $u_4 = \sqrt{3/4}\lambda^{-3/4}$ and the desired contradiction reduces to showing that

$$-2^{1/3} + \frac{8}{3}\left(1 - \sqrt{\frac{3}{4}}\lambda^{-3/4}\right)^{3/4} \geq 0.$$

or

$$\lambda^{-3/4} \leq \sqrt{\frac{4}{3}}\left(1 - \left(\frac{3\cdot 2^{1/3}}{8}\right)^{4/3}\right),$$

which is true for the range of λ under consideration. This shows the orbit cannot cross above the plane given by (13).

For $T_4 < s < T_7$, where $0 < u < u_4$ and the orbit cannot cross the plane

$$w = -2\sqrt{\frac{4}{3}}\frac{(1-u_4)^{3/4}}{u_4\lambda^{1/4}}u + \sqrt{\frac{4}{3}}\frac{(1-u_4)^{3/4}}{\lambda^{1/4}},$$

then all the more so the orbit must satisfy

$$w < \sqrt{\frac{4}{3}}\frac{(1-u)^{3/4}}{\lambda^{1/4}}.$$

This means that over the curve $u = 1 - \lambda v^2$ the Lyapunov function must be negative. This is in contradiction to $V_0(T_7) > 0$. Therefore, no internal tangency is possible.

Case 2. $u(T_2) \geq 1$.

Consider the "Lyapunov" function

$$V_\infty(u, w, z) = vw + \frac{\lambda u v^2}{2} - \frac{(u-1)^2}{2}$$

with derivative along an orbit

$$\dot{V}_\infty = w^2 + \frac{3\lambda v^3}{2}.$$

Under the assumption that there is an orbit from $(0, 0, w(0))$ to $(0, 0, w(T_7))$ with an internal tangency, then along this orbit

$$0 = V_\infty(0, 0, w(T_7)) - V_\infty(0, 0, w(0)) = \int_0^{T_7}\left(w(z)^2 + \frac{3\lambda v(z)^3}{2}\right) dz.$$

If we show that $\int_0^{T_7}(w^2 + (3\lambda v^3/2))\,dz > 0$, then we will have a contradiction. Thus, V_∞ does not have a (positive) sign-definite Lyapunov derivative, but along hypothesized orbits with an internal tangency its value is increasing, leading to a contradiction. The proof is stated in terms of a single loop from the initial condition on the w-axis back to the hypothetical tangency on the w-axis. Nevertheless, if there is a multiple loop through the plane $v = 0$ before returning to the tangency on the w-axis, the proof is the same by setting T_0 to the last time the orbit passes through the plane $v = 0$ with $v' > 0$ before the tangency at T_7. If the orbit starts from the v-axis, the proof is again the same, a comment that is important for the next section on the stress-free boundary conditions.

We will show the integral is positive by breaking the interval of integration into portions from 0 to T_2, and from T_2 to T_7,

$$\int_0^{T_7} w^2\,dz + \frac{3\lambda}{2}\left(\int_0^{T_2} v^3\,dz + \int_{T_2}^{T_7} v^3\,dz\right).$$

For $0 < s < T_2$ we have $u'(s) = v > 0$, so that it is possible to solve for s in terms of u and express v as a function of u on the interval $0 < u < u(T_2) \equiv u_2$, say the result is $v_1(u)$. Likewise on the interval $T_2 < s < T_7$, $u'(s) = v < 0$ and so we can express v

as a function of u for $0 < u < u_2$, calling the result $v_2(u)$. The integrals can then be expressed as

$$\int_0^{T_7} w^2\,dz + \frac{3\lambda}{2}\left(\int_0^{u_2} v_1(u)^2\,du - \int_0^{u_2} v_2(u)^2\,du\right).$$

We will show that $v_1(u) > -v_2(u)$, and so the integral is positive.

At $s = T_2$, $u(T_2) \geq 1$, $v(T_2) = 0$, and $w(T_2) < 0$ so that $v' = w < 0$. Thus it is possible to express s in terms of v and so to express u as a function of v in a neighborhood of $v = 0$ and $u = u_2$. In fact, $du/dv = v/w = 0$ at $s = T_2$ and

$$\left.\frac{d^2u}{dv^2}\right|_{T_2} = \frac{1}{w(T_2)} - \frac{v(T_2)w'(T_2)}{w(T_2)^3} = \frac{1}{w(T_2)}.$$

Note that $w'(T_2) = \lambda v^2(T_2) - \lambda u(T_2)w(T_2) + u(T_2) - 1 > 0$. For $T_2 - \epsilon < s < T_2$, $v > 0$ and $w < w(T_2) < 0$, Thus for $v > 0$, $d^2u/dv^2 = 1/w - vw'/w^3 > 1/w(T_2)$. Likewise for $v < 0$, $d^2u/dv^2 = 1/w - vw'/w^3 < 1/w(T_2)$. Then for $v > 0$, $u(v) > u_2 + (1/w(T_2))v^2 > u(-v)$. See Fig. 8. In terms of the functions $v_1(u)$ and $v_2(u)$, we have that $v_1(u) > -v_2(u)$ for $u - \delta < u < u_2$, for some small δ. We want to show that this inequality can be continued to the entire interval $0 < u < u_2$.

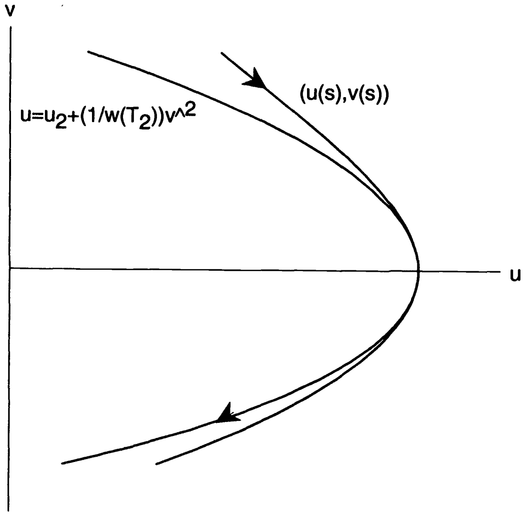


FIG. 8. *u as a function of v in a neighborhood of $v = 0$ and $u = u_2$.*

Consider $u^* = \max\{0 < u < u_2 : v_1(u) = -v_2(u)\}$. Let $v^* = v_1(u^*) = -v_2(u^*)$. There are two times $0 < s_1 < T_2 < s_2 < T_7$ such that $u(s_1) = u^* = u(s_2)$. Then at the point u^*,

$$\frac{w(s_2)}{v^*} = -\frac{w(s_2)}{v(s_2)} = -\left.\frac{dv_2}{du}\right|_{u^*} < \left.\frac{dv_1}{du}\right|_{u^*} = \frac{w(s_1)}{v(s_1)} = \frac{w(s_1)}{v^*}.$$

Therefore, $w(s_2) < w(s_1)$.

However, by the choice of $u^* = u(s_1) = u(s_2)$,

$$\begin{aligned}
&V_\infty(u(s_2), v(s_2), w(s_2)) - V_\infty(u(s_1), v(s_1), w(s_1))\\
&= \int_{s_1}^{s_2} w^2(z) + \frac{3\lambda v^3(z)}{2}\, dz\\
&= \int_{s_1}^{s_2} w^2(z)\, dz + \frac{3\lambda}{2}\left(\int_{u^*}^{u_2} v_1^2(u)\, du - \int_{u^*}^{u_2} v_2^2(u)\, du\right)\\
&= \int_{s_1}^{s_2} w^2(z)\, dz + \frac{3\lambda}{2}\left(\int_{u^*}^{u_2} (v_1^2(u) - v_2^2(u))\, du\right)\\
&> 0,
\end{aligned}$$

so that

$$v(s_1)w(s_1) + \frac{\lambda u(s_1)v^2(s_1)}{2} - \frac{(u(s_1)-1)^2}{2} < v(s_2)w(s_2) + \frac{\lambda u(s_2)v^2(s_2)}{2} - \frac{(u(s_2)-1)^2}{2}.$$

Because $u(s_1) = u(s_2)$ and $v(s_1) = -v(s_2)$, this reduces to $v(s_1)w(s_1) < v(s_2)w(s_2)$ or $w(s_1) < -w(s_2)$.

From the Lyapunov function V_0 we know that

$$\frac{w(s_2)^2}{2} - \frac{\lambda v(s_2)^3}{3} - (u(s_2)-1)v(s_2) < \frac{w(s_1)^2}{2} - \frac{\lambda v(s_1)^3}{3} - (u(s_1)-1)v(s_1)$$

or

$$w(s_2)^2 - w(s_1)^2 < -\frac{2\lambda v(s_1)^3}{3} - 2(u(s_1)-1)v(s_1).$$

From the previous paragraphs $w(s_2)^2 > w(s_1)^2$, and so the u^* and v^* cannot occur in the region $u \geq 1 - \lambda v^2/3$. Because $w(s_2) < w(s_1) < -w(s_2)$, we cannot have $w(s_2) > 0$.

Similar to what was shown in shown in Case 1, subcase (a), crossing the parabolic cylinder $u = 1 - \lambda v^2/3$ from $u > 1 - \lambda v^2/3$ to $u < 1 - \lambda v^2/3$ must take place with $-3\lambda/2 < w$. The value of V_0 at the crossing would then satisfy $V_0 < 9\lambda^2/8$. But the orbit cannot later recross the parabolic cylinder from inside to outside again, because as in Case 1, subcase (a), that would require $V_0 > 9\lambda^2/8$ in contradiction to the fact that V_0 is decreasing. Finally ihe orbit cannot cross the plane $w = 0$ with $u < 1-\lambda v^2/3$ because then $V_0 < 0$, in contradiction to the fact that $V_0 > 0$ on $0 < s < T_7$.

Finally then, the conclusion is established by ruling out all possibilities for the orbit to return to the positive w-axis. □

LEMMA 6. *Define* $\Sigma = \{(0,0,\beta) : 0 \leq \beta \leq A(\lambda) + 1\}$ *where A is the constant defined by Lemma* 4. *Then there is* β_0, $0 \leq \beta_0 \leq A(\lambda) + 1$ *such that the the orbit starting from* $(0,0,\beta_0)$ *remains in* W *for all* s; *that is,* $\Sigma \neq \Sigma^0$.

Proof. Consider an orbit starting from $(0, 0, A(\lambda) + 1)$. The orbit immediately enters the interior of the bounded region R, where $u' = v > 0$, and so u is increasing. The orbit must stay above the floor panel given in Lemma 4. In particular, the orbit cannot approach the equilibrium point $(1, 0, 0)$. The orbit also cannot exit through $w = 0$ because $w' > 0$ for $v > 1/\sqrt{\lambda}$ and $w = 0$. In summary, the conclusion of the technical Lemmas 3 and 4 is that the orbit must exit R through the back wall $u = 1$ into the octant $\mathcal{O}_1$. Lemma 1 of [13] reaches essentially the same conclusion, by analytic rather than geometric techniques. However, that lemma does not simultaneously establish the boundedness of the orbits.

It is easy to see by the vector field examination done previously on the boundaries of W that the orbit which begins at the origin $(0,0,0)$ of the phase space immediately enters the octant $\mathcal{O}_2$.

Now define the initial set Σ to be the connected interval on the w-axis from $(0,0,0)$ to $(0,0,w_0)$, where $w_0 = A(\lambda) + 1$ as mentioned previously. If all orbits starting on Σ leave the Wazewski set W then by Proposition 2, the exit time mapping from the segment Σ to the boundary of W is a homeomorphism. Yet Lemmas 3 and 4 show that one orbit exits W through one component of the exit set and goes into $\mathcal{O}_1$. Another orbit leaves W through the other component of the exit set into $\mathcal{O}_2$. This is a contradiction, so there is at least one orbit starting on Σ which does not leave the Wazewski set W. □

Finally, the orbit starting from $(0,0,\beta_0)$ identified in the previous Lemma must be the solution of (1) with the no-slip boundary conditions (2) . This follows from the next two lemmas.

LEMMA 7. *An orbit not entering $\mathcal{O}_1$ or $\mathcal{O}_2$ must remain bounded with $u > 0$.*

Proof. From its initial point $(0,0,\beta_0) \in \Sigma$ the orbit enters the region R given in Lemma 3. There is an upper bound on v for orbits in R, namely $v_0(1) + \Lambda$. The orbit cannot enter $\mathcal{O}_1$ and so must have $v' = w < 0$ if $u > 1$ and $v > 0$. Thus, the orbit has $v_0(1) + \Lambda$ as an upper bound for v.

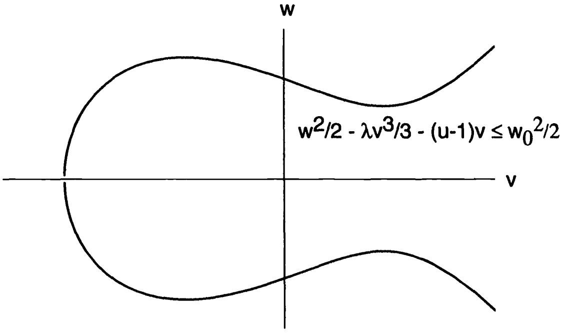


FIG. 9. *A cross section for $u \geq 0$ of the region containing the special orbit.*

Since the orbit does not enter $\mathcal{O}_2$, the Lyapunov function $V_0(u,v,w)$ is always decreasing. Therefore,

$$\frac{w^2}{2} - \frac{\lambda v^3}{3} - (u-1)v \leq \frac{w_0^2}{2},$$

where $w_0 = A(\lambda) + 1$ is the maximum w coordinate of the initial condition set Σ. Then for $u \geq 0$, $v \leq 0$ the orbits must lie in the region

$$-\sqrt{w_0^2 + 2(u-1)v + \frac{2\lambda v^3}{3}} \leq w \leq \sqrt{w_0^2 + 2(u-1)v + \frac{2\lambda v^3}{3}}.$$

See Fig. 9 for a sketch of the region. For each $u \geq 0$ the maximum extent in $v < 0$ occurs when

$$w_0^2 + 2(u-1)v + \frac{2\lambda v^3}{3} = 0.$$

A graph of u versus v for this relation shows that there is a minimum value of $v < 0$ satisfying this relationship; see Fig. 10. Combining this result with the bound for

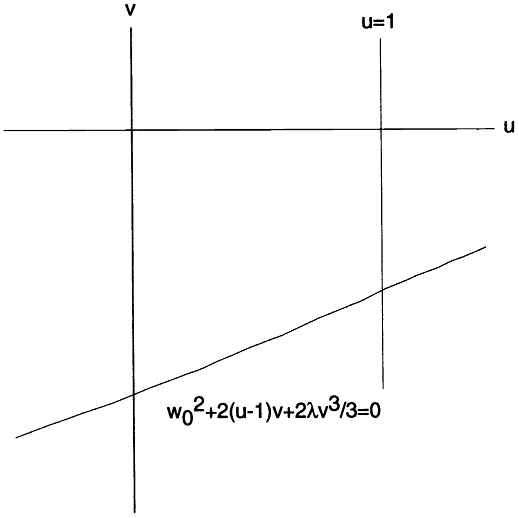


FIG. 10. *The graph of u versus v for* $w_0^2 + 2(u-1)v + 2\lambda v^3/3 = 0$.

$v > 0$ gives a global bound on the v coordinate of the special orbit. This argument is essentially the same as one already used in Lemma 5.

Since

$$\frac{w^2}{2} - \frac{\lambda v^3}{3} - (u-1)v \leq \frac{w_0^2}{2}$$

and v is bounded, if u is also bounded above, then w must be bounded.

Now for the Lyapunov function $V_\alpha(u,v,w) = w^2/2 - \lambda v^3/3 - (u-1)v + \alpha[vw + \lambda uv^2 - (u-1)^2/2]$ choose $\alpha < 0$ so small that $3\alpha\lambda v/2 - 1 < 0$ for all s. Then $\dot{V}_\alpha < 0$ and $V_\alpha(u,v,w)$ is decreasing along the special orbit. That is,

$$\frac{w^2}{2} - \frac{\lambda v^3}{3} - (u-1)v + \alpha\left[vw + \lambda uv^2 - \frac{(u-1)^2}{2}\right] \leq \frac{w_0^2}{2} - \frac{\alpha}{2}$$

for all s. We see that if w is bounded, then u must also be bounded.

Consequently, if the special orbit is unbounded, then u and w must be unbounded together. Since the orbit must always satisfy

$$\frac{w^2}{2} - \frac{\lambda v^3}{3} - (u-1)v \leq \frac{w_0^2}{2},$$

it is not possible for u and w to be simultaneously unbounded in $v < 0$. If u and w are unbounded together in the region $v > 0$ then ultimately $u > 1$ and $w < 0$ since the orbit cannot enter $\mathcal{O}_1$. But then $w' = \lambda v^2 - \lambda uw + (u-1) > 0$ and so w cannot decrease unboundedly.

This argument shows that the u and w coordinates must be bounded and so the special orbit is bounded. □

LEMMA 8. *The orbit starting from the initial point* $(0, 0, \beta_0)$ *must satisfy the boundary condition* $u(\infty) = 1$.

Proof. Since this special orbit must remain bounded with $u > 0$, consider the Lyapunov function

$$V_0(u,v,w) = \frac{w^2}{2} - \frac{\lambda v^3}{3} - (u-1)v.$$

It has already been noted that the derivative along an orbit $\dot{V}_0 = -\lambda uw^2 - v^2$ is negative so long as $u > 0$. By the LaSalle Invariance Principle, the ω-limit set is contained in the set where $\dot{V}_0 = 0$. Thus the ω-limit set of the orbit is an invariant set contained in the line $u > 0$, $w = 0$, $v = 0$ and so can only be the equilibrium point $(1, 0, 0)$. That is, the special orbit must satisfy the boundary condition $u(\infty) = 1$. □

This completes the proof of Theorem 1.

3. Existence of solutions to the stress-free problem. This section contains the proof of the following.

THEOREM 9. *For all $\lambda > 0$ there is at least one solution to the equation*

$$\phi''' + \lambda(\phi\phi'' - (\phi')^2) + 1 - \phi = 0,$$

satisfying the "stress-free" boundary conditions

$$\begin{aligned} \phi(0) &= 0, \\ \phi''(0) &= 0, \\ \phi(\infty) &= 1. \end{aligned}$$

The proof of this theorem is substantially the same as the proof of Theorem 1. The Wazewski set W is still appropriate and many of the same lemmas still apply. The main task is to identify the initial value set Σ and to establish that there are no internal tangencies.

LEMMA 10. *The orbit of the system* (7) *starting from the point* $(0, 1/\sqrt{\lambda}, 0)$ *on the positive v-axis leaves the set W and enters the octant $\mathcal{O}_1$.*

Proof. The initial tangent vector has $u'(0) = 1/\sqrt{\lambda}$, and $w'(0) = \lambda v^2 - \lambda uw + u - 1 = 0$ but with $w''(0) = 2\lambda vw - \lambda vw - \lambda uw' + v = 1/\sqrt{\lambda} > 0$ so that the orbit immediately enters the region R described in Lemma 3. Then the orbit has $w > 0$ and so v is increasing. This means in turn that the orbit stays in the region $v > 1/\sqrt{\lambda}$. Therefore, the orbit cannot exit the region R through the floor $w = 0$ since $w' > 0$ for $v > 1/\sqrt{\lambda}$. Furthermore, just as in Lemma 3, the orbit is bounded above in the v and the w coordinates. The u coordinate is increasing since $v > 0$. Thus the orbit must continue to increase until it exits the set W through the back wall $u = 1$ into the octant $\mathcal{O}_1$. □

LEMMA 11. *Define*

$$\Sigma = \left\{ (0, \beta, 0) : 0 \leq \beta \leq \frac{1}{\sqrt{\lambda}} \right\}.$$

Then orbits starting from the initial condition set Σ cannot have internal tangencies in the Wazewski set W.

Proof. The proof is the same as that of Lemma 5 since that proof does not depend on the initial conditions. □

The last argument to be filled in is the following.

LEMMA 12. *There is β_0, $0 \leq \beta_0 \leq 1/\sqrt{\lambda}$ such that the orbit starting from $(0, \beta_0, 0)$ remains in W for all s; that is, $\Sigma \neq \Sigma^0$.*

Proof. Consider the orbit starting from $(0, 1/\sqrt{\lambda}, 0)$. By the conclusion of technical Lemma 10 the orbit must exit R through the back wall $u = 1$ into the octant

$\mathcal{O}_1$. As noted before, the orbit from the origin $(0,0,0)$ of the phase space immediately enters the region $\mathcal{O}_2$.

Suppose all orbits starting on the connected interval Σ defined in Lemma 11 leave the Wazewski set W. Then by Proposition 2, the exit time mapping from Σ to the boundary of W is a homeomorphism. Yet one orbit exits W through one component of the exit set and enters $\mathcal{O}_1$. Another orbit leaves W through the other component of the exit set into $\mathcal{O}_2$. This is a contradiction, so there is at least one orbit starting on Σ which does not leave the Wazewski set W. □

The orbit starting from $(0,\beta_0,0)$ identified in the previous Lemma 12 must be bounded with $u > 0$ as shown by the proof of Lemma 7. The orbit starting from the initial point $(0,\beta_0,0)$ satisfies the boundary condition $u(\infty) = 1$ by the same argument as in Lemma 8. This finishes the proof of Theorem 9.

Acknowledgment. I would like to thank Professor Lloyd Jackson for helpful conversations during the course of this work.

REFERENCES

[1] L. R. ANDERSON, *Weighted Liapunov functions for a class of third-order autonomous differential equations*, Quart. Appl. Math., 45 (1987), pp. 481–486.

[2] C. CONLEY, *Isolated Invariant Sets and the Morse Index*, CBMS Regional Conference Series, No. 38, American Mathematical Society, Providence, RI, 1978.

[3] S. R. DUNBAR, *Travelling wave solutions of diffusive Lotka–Volterra equations*, J. Math. Biol., 17 (1983), pp. 11–32.

[4] ———, *Travelling wave solutions of diffusive Lotka-Volterra equations: A heteroclinic connection in R^4*, Trans. Amer. Math. Soc. 286 (1984), pp. 557–594.

[5] F. GANTMACHER, *The Theory of Matrices, Vol.* 2, Chelsea Publishing, New York, NY, 1964.

[6] I. S. GRADSHTEYN AND I. M. RYZHIK, *Table of Integrals, Series, and Products*, Academic Press, New York, London, 1980.

[7] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, Baltimore, MD, 1973.

[8] S. P. HASTINGS, *On the existence of homoclinic and periodic orbits for the Fitzhugh–Nagumo equations*, Quart. J. Math. Oxford Ser. (2), 27 (1976), pp. 123–134.

[9] S. P. HASTINGS AND W. C. TROY, *Oscillating solutions of the Falkner–Skann equation for positive β*, J. Differential Equations, 71 (1988), pp. 123–144.

[10] P. J. HOLMES AND D. S. STEWART, *The existence of one dimensional steady detonation waves in a simple model problem*, Stud. Appl. Math., 66 (1982), pp. 121–143.

[11] G. R. IERLEY AND O. G. RUEHR, *Analytical and numerical solutions of a nonlinear boundary-layer problem*, Stud. Appl. Math., 75 (1986), pp. 1–36.

[12] C. JONES AND T. KÜPPER, *On the infinitely many solutions of a semilinear elliptic equation*, SIAM J. Math. Anal., 17 (1986), pp. 803–835.

[13] W. C. TROY, *Solutions of a nonlinear boundary layer problem arising in physical oceanography*, preprint.

[14] ———, *The existence of travelling wave front solutions of a model of the Belousov–Zhabotinski chemical reaction*, J. Differential Equations, 36 (1980), pp. 89–98.

SIAM J. MATH. ANAL.
Vol. 24, No. 2, pp. 466-479, March 1993

BOUNDARY AND INTERIOR SPIKE LAYER FORMATION FOR AN ELLIPTIC EQUATION WITH SYMMETRY*

ARNALDO S. DO NASCIMENTO†

Abstract. Consider the problem $\nabla[k(\|x\|)\nabla U_\lambda]+\lambda f(U_\lambda)=0$, $x\in B$: unit ball in R^n, $\nabla U_\lambda\cdot\hat{n}/_{\partial B}=0$. Conditions are given on k and f that guarantee the formation of boundary layer and also of a spike layer around the origin as $\lambda\to\infty$ for some particular radially symmetric solutions to the above problem. In the first case, the nodal curve of the solution approaches ∂B as $\lambda\to\infty$, while in the second one it shrinks toward the origin as $\lambda\to\infty$. Range location for the value of the solution on ∂B, in the first case, and at the origin in the second one, is given for λ sufficiently large. Only radially symmetric solutions are considered since they are the only ones that can be stable equilibria of the corresponding parabolic equation.

Key words. convergence in measure, nodal curves, maximum principle, radially symmetric solution

AMS(MOS) subject classifications. 35B25, 35B40, 34E15

1. Introduction. Consider the following nonlinear eigenvalue problem:

$$\begin{aligned}&\nabla[k(\|x\|)\nabla U_\lambda]+\lambda f(U_\lambda)=0, \qquad x\in B,\\ &\nabla U_\lambda\cdot\hat{n}/_{\partial B}=0,\end{aligned}\tag{1.1}$$

where B is the unit ball in R^n, $\hat{n}$ the unit vector orthogonal to ∂B, $\|\cdot\|$ the Euclidian norm, $\lambda\in(0,\infty)$, $k(\|x\|)$ a positive function in $C^2(\bar{B},R)$, and $f\in C^2(R,R)$ satisfies

$$\exists a, b\in R, b<0<a: f(b)=f(0)=f(a)=0, \qquad f'(0)>0, \tag{h$_1$}$$

$$\operatorname{sgn} f''(U)=-\operatorname{sgn} U \quad \forall U\in R,\ U\neq 0. \tag{h$_2$}$$

To simplify notation we shall drop the subscript λ in the notation of the solution U_λ and consider $n=2$.

We restrict our study to the radially symmetric solutions to (1.1) due to the fact that if the corresponding parabolic problem is considered, that is,

$$\begin{aligned}&\frac{\partial U}{\partial t}=\nabla[k(\|x\|)\nabla U]+\lambda f(U), \quad (t,x)\in(0,\infty)\times B,\\ &\nabla U\cdot\hat{n}/_{\partial B}=0,\end{aligned}\tag{1.2}$$

then any stable equilibrium solution to (1.2) in $W^{1,2}(B)$ must exhibit such symmetry. Conversely, if $W_r^{1,2}(B)$ stands for the space of radially symmetric functions of $W^{1,2}(B)$, then any radially symmetric equilibrium of (1.2), which is stable in $W_r^{1,2}(B)$, turns out to be stable in the larger space $W^{1,2}(B)$, too. Also it is well known that if $k(\|x\|)=$ constant, then any stable equilibrium solution of (1.2) must be a constant function. This is no longer the case if $k(\|x\|)$ is allowed to vary in a suitable manner. These results will appear in a forthcoming paper.

Roughly speaking, by boundary layer formation we mean that a radial solution $U(x,\lambda)$ to (1.1) exhibits the following geometrical feature: for any R, $0<R<1$, $r=\|x\|$, $U(r,\lambda)$ converges uniformly to a constant solution on $[0,R]$, as $\lambda\to\infty$ and $U(r,\lambda)$ varies abruptly in $[R,1]$ so that $|U_r(r,\lambda)|$ assumes arbitrarily large values in $[R,1]$ as $\lambda\to\infty$.

* Received by the editors June 24, 1991; accepted for publication (in revised form) May 28, 1992.

† Universidade Federal de São Carlos-DM, 13560—São Carlos, SP, Brasil.

The very same interpretation can be given for interior spike layer formation by replacing the above interval by $[0, R]$.

Borrowing the terminology used in the literature for reaction-diffusion equations, we herein refer to k as the diffusion function and to f as the reaction term.

The following result has been proved in [1]. Let $\tilde{\lambda}_n$, $n \geqq 0$, be the nth eigenvalue of the linearized problem ($\tilde{\lambda}_0 = 0$):

$$(1.3)\qquad \begin{aligned} &(k(r)U_r)_r + k(r)\frac{U_r}{r} + \lambda f'(0)U = 0, \qquad r \in (0, 1), \\ &\qquad\qquad\qquad U_r(0) = 0, \qquad U_r(1) = 0. \end{aligned}$$

Suppose that f satisfies (h_1) and (h_2) and the diffusion function k satisfies

$$(1.4)\qquad r^2(k^{1/2})_{rr} + r(k^{1/2})_r \leqq k^{1/2}, \qquad 0 < r < 1.$$

Note that $k \equiv$ const. satisfies the inequality.

If $\lambda \in (\tilde{\lambda}_n, \tilde{\lambda}_{n+1})$, $n \geqq 1$, then there are exactly $(2n+3)$ radially symmetric solutions of (1.1), three of them being the constant solutions 0, a, and b. For $\lambda \in (0, \tilde{\lambda}_1)$, the only solutions are the constant ones: 0, a, and b.

If C_k, $k \geqq 1$, denotes the branch of radially symmetric solutions of (1.1) bifurcating from the zero solution at $\lambda = \tilde{\lambda}_k$, then C_k is defined for all $\lambda \geqq \tilde{\lambda}_k$, and the solutions in it, denoted by $\phi_k(r, \lambda)$, $0 \leqq r \leqq 1$, are characterized by having exactly k simple zeros in $(0, 1)$ and satisfy $b < \phi_k(r, \lambda) < a$, $r \in [0, 1]$.

Moreover, $C_k = C_k^+ \cup C_k^-$, where $\phi_k \in C_k^+$ if $\phi_k(0, \lambda) > 0$, $\phi_k \in C_k^-$ if $\phi_k(0, \lambda) < 0$.

Throughout this paper by $\phi_n^+(r, \lambda)[\phi_n^-(r, \lambda)]$, $n \geqq 1$, we mean a solution to (1.1) in $C_n^+[C_n^-]$, $\lambda > \tilde{\lambda}_n$, and $r_1(\lambda), \dots, r_n(\lambda)$, where $0 < r_1(\lambda) < \cdots < r_n(\lambda) < 1$ stand for the zeros of $\phi_n^+(r, \lambda)[\phi_n^-(r, \lambda)]$ in $(0, 1)$. They are simple zeros and, in particular, $(d/dr)\,\phi_1^+(r, \lambda) < 0$ $[(d/dr)\phi_1^-(r, \lambda) > 0]$ in $(0, 1)$.

In other words, the above results yield the global bifurcation diagram for the radially symmetric solutions to (1.1). We are concerned here with the asymptotic behavior of the solution $\phi_1^+(., \lambda)[\phi_1^-(., \lambda)]$ as we follow the bifurcation branch $C_1^+[C_1^-]$.

We prove that if, besides the assumptions (h_1) and (h_2), f and k are related to each other according to

$$(h_3)\qquad k_m \int_0^b f > k_M \int_0^a f,$$

where $k_m = \min_{0 \leqq r \leqq 1} k(r)$ and $k_M = \max_{0 \leqq r \leqq 1} k(r)$, then $r_1(\lambda) \to 1$ as $\lambda \to \infty$. Also $\phi_1^+(r, \lambda)$ converges uniformly as $\lambda \to \infty$ to the constant solution $U(r) = a$ in $[0, \bar{r}]$ for any $\bar{r}$ such that $0 < \bar{r} < 1$.

It follows from (h_3) that there are number α and β, $b < \beta \leqq \alpha < 0$ satisfying

$$\frac{k_M}{k_m}\int_0^a f = \int_0^\beta f \quad \text{and} \quad \frac{k_m}{k_M}\int_0^a f = \int_0^\alpha f.$$

Then we also prove that $\beta \leqq \lim_{n\to\infty} \inf \phi_n^+(1, \lambda_n) \leqq \lim_{n\to\infty} \sup \phi_n^+(1, \lambda) \leqq \alpha < 0$ for any sequence $\{\lambda_n\}_{n \in N}$ such that $\lambda_n \to \infty$ as $n \to \infty$. Note that the closer k_m is taken to k_M, the nearer α is to β. See Fig. 1 below. In particular, for the case of constant diffusion, that is, $k(r) \equiv$ const., then $\lim_{\lambda\to\infty} \phi_1^+(1, \lambda) = c$ where c is the unique negative number, $b < c < 0$ so that

$$\int_0^a f = \int_0^c f, \quad \text{where } b < c < 0 < a.$$

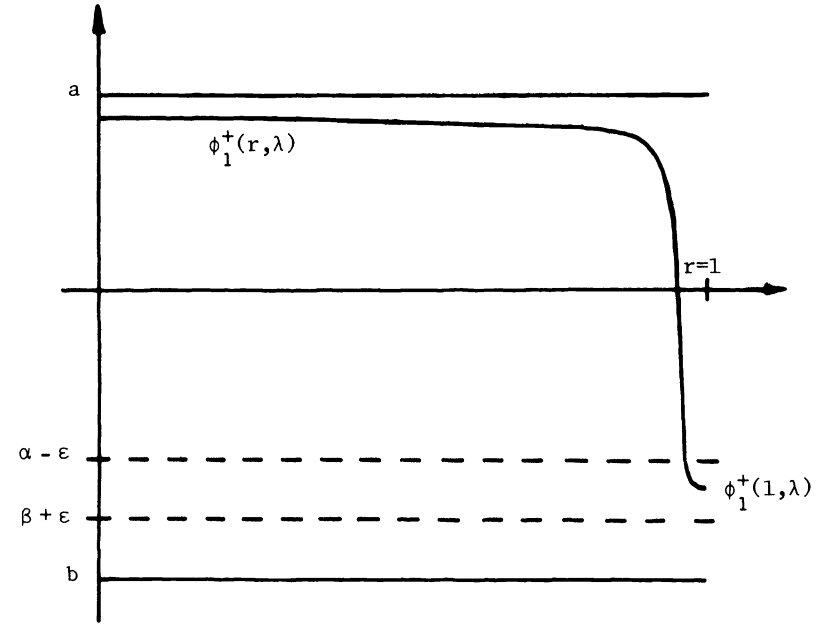


FIG. 1

Condition (h_3) means that the area of the negative bump of f, namely, $\int_0^b f$, is bigger than the area of the positive bump of f, namely, $\int_0^a f$, multiplied by (k_M/k_m). In particular, altogether it implies that

$$\int_0^b f > \left(\frac{k_M}{k_m}\right)\int_0^a f = \int_0^\beta f = \left(\frac{k_M}{k_m}\right)\int_0^\alpha f.$$

Using hypotheses (h_1), (h_2), and (h_3), similar results hold for solutions $\phi_1^-(r,\lambda)$, as we follow the bifurcation half branch C_1^-, except that now $r_1(\lambda)\to 0$ as $\lambda\to\infty$. Also it follows that for any $\varepsilon>0$, ε: small, there is $\Lambda=\Lambda(\varepsilon)$ such that $b<\beta+\varepsilon<\phi_1^+(1,\lambda)<\alpha+\varepsilon<0$ for $\lambda\geqq\Lambda$.

Note that by setting $\lambda=1/\varepsilon$, then as $\lambda\to\infty$, (1.1) can be viewed as the following singularly perturbed semilinear elliptic problem:

$$\varepsilon\nabla[k(\|x\|)\nabla U_\varepsilon]+f(U_\varepsilon)=0,$$
$$U_\varepsilon.\hat{n}/_{\partial B}=0.$$

There is a vast literature in singularly perturbed problems, and among others the reader is referred to [2], [3], [4], [5], [8], [10], and [11].

In this paper the following should be seen as the major contributions: the emphasis on the relationship between the diffusion function k and the reaction term f in determining the asymptotic behavior of the nodal curves of $\phi_1^+(.,\lambda)$ and $\phi_1^-(.,\lambda)$, the formation of layer along with its location, and finally the asymptotic behavior of the solutions on the layer's locus.

Herein, once a solution $\phi_n^+(.,\lambda)\in C_n^+$ is taken, we follow the bifurcation branch, and the results are accomplished by using comparison techniques based on maximum principles, convergence in measure to the equilibrium solutions and equalities that relate the areas of the negative and positive bumps of f, up to the end points of the first half loops of the solution $\phi_1^+(r,\lambda)$ in the phase plane and the maximum and minimum attained by $k(r)$, $r\in[0,1]$.

Condition (1.4) above was assumed in [1] to guarantee that each C_n^+, C_n^-, $n \geqq 1$, was monotone in λ and did not present secondary bifurcation. However, for the purpose of this work (1.4) need not be assumed. Indeed, all we need here is to have each C_n^+, C_n^-, $n \geqq 1$ defined for all λ, $\lambda > \tilde{\lambda}$, and the nodal properties which follow from the work of Rabinowitz [6] and from the Sturm-Liouville theory, respectively.

2. The case $\phi_1^+(.,\lambda) \in C_1^+$: boundary layer. In this section we only deal with the solution $\phi_1^+(.,\lambda)$, and, therefore, for simplicity in notation we drop the superscript +, thus just writing $\phi_1(.,\lambda)$.

In view of the previous considerations, instead of problem (1.1), it suffices to consider

$$\begin{gathered}(k(r)U_r)_r + k(r)\frac{U_r}{r} + \lambda f(U) = 0, \qquad 0 < r < 1,\\ U_r(0) = U_r(1) = 0.\end{gathered} \tag{2.1}$$

LEMMA 2.1. *Let* $k_m = \min_{0 \leqq r \leqq 1} k(r)$, $k_M = \max_{0 \leqq r \leqq 1} k(r)$ *and suppose that* f *satisfies* (h_1), (h_2), *and* (h_3). *Then it holds that*

$$k_m \int_0^{\phi_1(1,\lambda)} f < k_M \int_0^{\phi_1(0,\lambda)} f$$

for any $\lambda > \tilde{\lambda}_1$, *where* $\tilde{\lambda}_1$ *is the first positive eigenvalue of problem* (1.2).

Proof. The solution $\phi_1(r,\lambda)$ satisfies

$$\frac{d}{dr}\left[\frac{(k\phi_{1,r})^2}{2} + \int_0^r \frac{(k\phi_{1,r})^2}{s}\,ds + \lambda \int_0^r k(s)\frac{d}{ds}F(\phi_1(s,\lambda))\,ds\right] = 0$$

for $r \in (0,1)$, where $F(\phi) = \int_0^\phi f$.

Therefore, the expression between brackets is constant in [0, 1], and taking $r = 0$, we conclude that this constant is zero. Since $\phi_{1,r}(1,\lambda) = 0$, we obtain

$$\int_0^1 \frac{(k(s)\phi_{1,s})^2}{s}\,ds + \lambda \int_0^1 k(s)\frac{d}{ds}F(\phi_1(s,\lambda))\,ds = 0,$$

which, by its turn, implies that

$$\left(\int_0^{r_1(\lambda)} kf(\phi_1)\phi_{1,s}\,ds + \int_{r_1(\lambda)}^1 kf(\phi_1)\phi_{1,s}\,ds\right) < 0,$$

for all $\lambda > \tilde{\lambda}_1$.

On $[0, r_1(\lambda)]$ it holds that $f(\phi_1(r,\lambda)) \geqq 0$, $\phi_{1,r} \leqq 0$ and on $[r_1(\lambda), 1]$, $f(\phi_1(r,\lambda)) \leqq 0$, and $\phi_{1,r} \leqq 0$.

Hence

$$0 < \int_{r_1(\lambda)}^1 kf(\phi_1)\phi_{1,s}\,ds < -\int_0^{r_1(\lambda)} kf(\phi_1)\phi_{1,s}\,ds,$$

and with k_m and k_M as defined above:

$$0 < k_m \int_{r_1(\lambda)}^1 \frac{d}{ds}F(\phi_1(s,\lambda))\,ds < -k_M \int_0^{r_1(\lambda)} \frac{d}{ds}F(\phi_1(s,\lambda))\,ds.$$

But $F(\phi_1(r_1,\lambda)) = 0$, and then $k_m F(\phi_1(1,\lambda)) < k_M F(\phi_1(0,\lambda))$, that is,

$$k_m \int_0^{\phi_1(1,\lambda)} f < k_M \int_0^{\phi_1(0,\lambda)} f,$$

and the lemma is proved.

In what follows, Lemma 2.1 will be useful in the sense that if $\{\lambda_n\}_{n\in N}$, $\lambda_n\to\infty$ as $n\to\infty$ is such that $\phi_1(1,\lambda_n)$ converges as $n\to\infty$, then we easily conclude that $b<\lim_{n\to\infty}\phi(1,\lambda_n)\leqq 0$. This yields information about the asymptotic behavior of $r_1(\lambda_n)$ for if $r_1(\lambda_n)\not\to 1$, as $n\to\infty$, by constructing a supersolution in a convenient interval we are able to prove that $\lim_{n\to\infty}\phi(1,\lambda_n)=b$.

We will work now toward proving that $r_1(\lambda)\to 1$, as $\lambda\to\infty$ and to this end let us consider the following linear eigenvalue problem:

$$(2.2)\qquad \begin{aligned}&(k\psi_r)_r+k\frac{\psi_r}{r}+\mu f'(0)\psi=0, \qquad R<r<1,\\ &\psi(R)=\psi(1)=0,\end{aligned}$$

where $0<R<1$. It is well known that the eigenfunction $J(r)$ corresponding to the first eigenvalue μ_0 of (2.2) is of one sign in $(R,1)$. By multiplying by a constant, if necessary, we can suppose $b<J(r)<0$ in $(R,1)$.

LEMMA 2.2. *Suppose that there is R, $0<R<1$ and a sequence $\{\lambda_n\}_{n\in N}$, $\lambda_n>\tilde\lambda_1$, $\lambda_n\to\infty$ as $n\to\infty$ such that $r_1(\lambda_n)$; the unique zero of $\phi_1(r,\lambda_n)$ in $(0,1)$ satisfies $0<r_1(\lambda_n)\leqq R$ for any n. Let μ_0 and $J(r)$ be the principal eigenvalue and eigenfunction of* (2.2), *as above. Then $J(r)\geqq\phi_1(r,\lambda_n)$ in $[R,1]$ for n sufficiently large.*

Proof. On $[R,1]$ we have $\phi_1(r,\lambda_n)\leqq 0$ and $\phi_{1,r}(1,\lambda_n)=0$. Moreover, J can be taken as negative in $(R,1)$.

Let us suppose by contradiction that for any Λ, there is n such that $\lambda_n\geqq\Lambda$ and $\tilde r\in[R,1]$ such that $0>\phi_1(\tilde r,\lambda_n)>J(\tilde r)$. By continuity there are numbers $\tilde r_1(\lambda_n)$ and $\tilde r_2(\lambda_n)$, $\tilde r_1(\lambda_n)\leqq\tilde r(\lambda_n)<\tilde r_2(\lambda_n)<1$ such that $J(\tilde r_1)=\phi_1(\tilde r_1,\lambda_n)$, $J(\tilde r_2)=\phi_1(\tilde r_2,\lambda_n)$, and $\phi_1(r,\lambda_n)>J(r)$ in $(\tilde r_1,\tilde r_2)$.

There are two cases to consider.

Case 1. $r_1(\lambda_n)<R$ for infinitely many n.

In this case, as a consequence, $R<\tilde r_1(\lambda_n)<\tilde r_2(\lambda_n)<1$. In order to resort to a comparison technique, we set

$$\xi(r)=\frac{J(r)}{\phi_1(r,\lambda_n)},$$

which satisfies

$$\xi_{rr}+\left[\frac{k_r}{k}+\frac{1}{r}+\frac{2\phi_{1,r}}{\phi_1}\right]\xi_r+\frac{1}{k}\left[\mu_0 f'(0)-\lambda_n\frac{f(\phi_1(r,\lambda_n))}{\phi_1(r,\lambda_n)}\right]\xi=0$$

for $\tilde r_1(\lambda_n)<r<\tilde r_2(\lambda_n)$ and $\xi(\tilde r_1)=\xi(\tilde r_2)=1$.

Note that by virtue of hypotheses (h_1) and (h_2) in $(\tilde r_1,\tilde r_2)$ it holds that

$$f'(0)\geqq\frac{f(\phi_1(r,\lambda_n))}{\phi_1(r,\lambda_n)}\geqq\frac{f(J(r))}{J(r)}\geqq\frac{f(J_m)}{J_m}>0,$$

where $J_m=\min\{J(r),\ \tilde r_1(\lambda)\leqq r\leqq\tilde r_2(\lambda)\}$.

The assumption that for any Λ there is $\lambda_n\geqq\Lambda$ and $\tilde r$, $R<\tilde r<1$, such that $\phi_1(\tilde r,\lambda_n)>J(\tilde r)$ prevents $\phi_1(r,\lambda_n)$ from approaching the equilibrium solution $b\equiv$ const. for λ_n: large, and this by its turn yields a lower bound for $[f(\phi_1(r,\lambda_n))/\phi_1(r,\lambda_n)]$. Therefore, by taking Λ large enough it is possible to make the coefficient of ξ in equation above negative in $(\tilde r_1,\tilde r_2)$.

Summing up we have $\xi(\tilde r_1)=\xi(\tilde r_2)=1$ and $\xi(r)>0$ in $(\tilde r_1,\tilde r_2)$. In particular, ξ assumes its maximum in $(\tilde r_1,\tilde r_2)$ since $\xi(\tilde r)>1$. But this is impossible according to the remarks above and a well-known maximum principle. See [9].

Case 2. $r_1(\lambda_n) = R$ for infinitely many n.

In this case we have $R = \tilde{r}_1(\lambda_n) < \tilde{r}_2(\lambda_n) < 1$ and ξ would still satisfy (2.3) with different boundary conditions, namely, $\xi(R) > 0$, $\xi(1) = 0$, $\xi_r(R) = 0$, and $\xi > 0$ in $(R, 1)$.

The computation $\xi_r(R) = 0$ can be accomplished by applying l'Hôpital's rule twice and using (2.2) and (2.1).

Now an application of the maximum principle referred to above yields a contradiction, and the lemma is proved.

LEMMA 2.3. *Under the hypotheses and notation of Lemma* 2.2 *it holds that*

$$\lim_{n\to\infty} \int_R^1 J(r) f(\phi_1(r, \lambda)) r\, dr = 0.$$

Proof. Since $\phi_1(r, \lambda_n)$ is an equilibrium solution of (2.1), with the notation set forth there, but by dropping the index n for simplicity we have

$$J(r)(rk(r)\phi_{1,r})_r = -\lambda r f(\phi_1(r, \lambda))J(r), \qquad R \leqq r \leqq 1.$$

But $(rk\phi_{1,r}J)_r = J(rk\phi_{1,r})_r + rJ_r\phi_{1,r}k(r)$, and, therefore, $-\lambda r f(\phi_1)J(r) = (rk\phi_{1,r}J)_r - rJ_r\phi_{1,r}k$. Again $(rk\phi_1 J_r)_r = r\phi_{1,r}kJ_r + \phi_1(rkJ_r)_r = r\phi_{1,r}J_r k - f'(0)\mu_0 r\phi_1 J$, since J satisfies problem (2.2).

Henceforth by combining the last two equalities, $(rJk\phi_{1,r})_r - (rk\phi_1 J_r)_r - f'(0)\mu_0 r\phi_1 J = -\lambda rJ f(\phi_1)$.

Integrating the above equality from R to 1 and bearing in mind that $J(R) = 0$, $\phi_{1,r}(1, \lambda) = 0$ we obtain

$$[-rk(r)\phi_1(r, \lambda)J_r(r)]/{}_R^1 - f'(0)\mu_0 \int_R^1 J(r)\phi_1(r, \lambda) r\, dr = -\lambda \int_R^1 J(r) f[\phi_1(r, \lambda)] r\, dr.$$

In $[R, 1]$, $b < \phi_1(r, \lambda) < 0$, for any $\lambda \geqq \tilde{\lambda}_1$ and since J does not depend on λ we conclude that

$$\lim_{\lambda\to\infty} \int_R^1 J(r) f(\phi_1(r, \lambda)) r\, dr = 0.$$

LEMMA 2.4. *Under the hypotheses of Lemma* 2.2 *and with notation set forth there it holds that* $\phi_1(r, \lambda_n) \to b$ *in* $[R, 1]$, *in measure as* $n \to \infty$. *In particular,* $\phi_1(1, \lambda_n) \to b$ *as* $n \to \infty$.

Proof. Again for simplicity we just write λ for λ_n.

Roughly speaking, the idea of the proof consists of first concluding that $\phi_1(., \lambda)$ approaches in measure a zero of f, which can be seen from the previous lemma and second by using Lemma 2.2, which yields a super solution to the equilibria of problem (2.1) in $[R, 1]$ and, consequently, prevents $\phi_1(., \lambda)$ from approaching the zero solution in $[R, 1]$.

There remains only the constant solution b for $\phi_1(., \lambda)$ to converge to, as $\lambda \to \infty$. A similar idea has previously been used by de Figueiredo in [7].

To this end let us take $\varepsilon > 0$ arbitrarily small and set $I_\varepsilon = \{r \in [R, 1]: R + \varepsilon \leqq r \leqq 1 - \varepsilon\}$. Recall that by Lemma 2.2 $0 \geqq J(r) \geqq \phi_1(r, \lambda) > b$, in I_ε for λ sufficiently large.

Therefore, there is $m(\varepsilon) < 0$, $m(\varepsilon) = \sup\{J(r), r \in I_\varepsilon\}$ so that

$$\int_{I_\varepsilon} J(r) f(\phi_1(r, \lambda)) r\, dr \geqq m(\varepsilon) \int_{I_\varepsilon} f(\phi_1(r, \lambda)) r\, dr > 0$$

for any λ sufficiently large.

Suppose now that there is a $\delta > 0$ and a subsequence $\{\lambda_n\}$, $\lambda_n \to \infty$ as $n \to \infty$, such that the Lebesgue measure m^* of the sets $I_{\varepsilon,n} = \{r \in I_\varepsilon m(\varepsilon) \geqq \phi_1(r, \lambda_n) \geqq b + \varepsilon\}$ satisfies $m^*(I_{\varepsilon,n}) \geqq \delta$ for any n.

It follows from the hypotheses on f, namely, (f_1) and (f_2) that there is negative constant f_m so that $f(\phi_1(r, \lambda_n)) \leqq f_m$ in $I_{\varepsilon,n}$.

Therefore,

$$m(\varepsilon) \int_{I_\varepsilon} f(\phi_1(r, \lambda_n)) r\, dr \geqq m(\varepsilon) \int_{I_{\varepsilon,n}} f(\phi_1(r, \lambda_n)) r\, dr$$

$$\geqq m(\varepsilon) f_m m^*(I_{\varepsilon,n}) \geqq m(\varepsilon) f_m \delta > 0 \quad \text{for any } n.$$

But this is a contradiction since from the first chain of inequalities in this proof and Lemma 2.2 it follows that

$$\lim_{n\to\infty} \int_{I_\varepsilon} f(\phi_1(r, \lambda_n)) r\, dr = 0.$$

Therefore, for any $\varepsilon > 0$, $\lim_{n\to\infty} m^*(I_{\varepsilon,n}) = 0$, and our claim follows. Also $\phi_1(1, \lambda) \to b$, as $\lambda \to \infty$, since $\phi_{1,r} < 0$ on $[R, 1)$.

The sequence of previous results culminate with the following theorem.

THEOREM 2.5. *If* $k_m \int_0^b f > k_M \int_0^a f$ *holds, with* k_m *and* K_M *as before, then* $r_1(\lambda)$, *the unique zero of the solution* $\phi_1(r, \lambda)$ *in* $[0, 1]$, *satisfies* $\lim_{\lambda\to\infty} r_1(\lambda) = 1$.

Proof. Suppose by contradiction that there is a subsequence $\{\lambda_n\}$ and R, $0 < R < 1$ such that $0 < r_1(\lambda_n) \leqq R$.

The same notation of Lemma 2.2 is being used for simplicity, and hopefully this will not cause any confusion.

By Lemma 2.1,

$$k_m \int_0^{\phi_1(1,\lambda_n)} f < k_M \int_0^{\phi_1(0,\lambda_n)} f, \qquad \lambda > \tilde{\lambda}_1,$$

and $\lim_{n\to\infty} \phi_1(1, \lambda_n) = b$ by Lemma 5.4. Therefore, taking the limit, as $n \to \infty$, we obtain

$$k_m \int_0^b f \leqq \lim_{n\to\infty} k_M \int_0^{\phi_1(0,\lambda_n)} f \leqq k_M \int_0^a f,$$

which is against our assumption.

LEMMA 2.6. *Under the hypotheses* (h_1), (h_2), *and* (h_3) *it holds that* $\phi_1(r, \lambda) \to a$, *in measure, for* $0 \leqq r \leqq 1$, *as* $\lambda \to \infty$. *In particular,* $\phi_1(r, \lambda) \to a$, *uniformly, in* $[0, \bar{r}]$, *as* $\lambda \to \infty$ *for any* $\bar{r}$ *such that* $0 < \bar{r} < 1$.

Proof. Let us take $\varepsilon > 0$. Consider $r_\varepsilon = 1 - (\varepsilon/2)$ and the following eigenvalue problem:

$$(k\psi_r)_r + k\frac{\psi_r}{r} + \mu f'(0)\psi = 0, \qquad 0 < r < r_\varepsilon,$$

$$\psi_r(0) = \psi(r_\varepsilon) = 0.$$

Let μ_0 be the first positive eigenvalue of the above problem and J its corresponding eigenfunction. We can take $J(r) > 0$ in $[0, r_\varepsilon)$.

Hereafter the proof parallels the proofs of the foregoing lemmas, and, therefore, we go rather fast, omitting some details.

Note that since $r_1(\lambda)$, the unique zero of $\phi_1(r, \lambda)$ in $(0, 1)$ satisfies $r_1(\lambda) \to 1$, as $\lambda \to \infty$, we can suppose that for λ: large enough, $r_\varepsilon < r_1(\lambda) < 1$.

Proceeding as in the proof of Lemma 2.2 it can be proved that for λ sufficiently large, $J(r) \leqq \phi_1(r, \lambda)$ in $[0, r_\varepsilon]$.

The very same computations done in the proof of Lemma 2.3 yield

$$\lim_{\lambda\to\infty}\int_0^{r_\varepsilon} J(r)f(\phi_1(r,\lambda))r\,dr=0. \tag{2.3}$$

We now use an argument similar to that used in the proof of Lemma 2.4 to conclude, in view of the above equality that $\phi_1(r,\lambda)\to a$, in $[0,r_\varepsilon]$, as $\lambda\to\infty$, in measure. Since $\phi_1(r,\lambda)$ is a decreasing function in $[0,r_\varepsilon]$, the Lebesgue measure m^* of the set $J_\varepsilon(\lambda)=\{r\in[0,r_\varepsilon]\colon 0\leqq\phi_1(r,\lambda)<a-\varepsilon\}$ satisfies $\lim_{\lambda\to\infty} m^*J_\varepsilon(\lambda)=0$, and so for λ: large enough we have $J_\varepsilon(\lambda)\subset[1-\varepsilon,r_\varepsilon]$. Hence for λ: large enough $m^*\{r\in[0,1]\colon b<\phi_1(r,\lambda)<a-\varepsilon\}<\varepsilon$. As for the uniform convergence of $\phi_1(r,\lambda)$ to a, in $[0,\bar r]$, as $\lambda\to\infty$, for any $\bar r\in(0,1)$ it can be proved by contradiction once we realize that $\phi_1(r,\lambda)$ is a decreasing function in $[0,1]$ and use (2.3) above along with Theorem 2.5.

The claim follows.

LEMMA 2.7. *Assuming the existence of* $\phi_{1,\infty}(1)=\lim_{n\to\infty}\phi_1(1,\lambda_n)$, *it holds that* $b<\phi_{1,\infty}(1)\leqq 0$, *where* $\lambda_n\to\infty$ *as* $n\to\infty$.

Proof. Taking the limit as $n\to\infty$ in the inequality derived in Lemma 2.1 and using hypothesis (h_3) we obtain

$$k_m\int_0^{\phi_{1,\infty}(1)} f\leqq k_M\int_0^a f<k_m\int_0^b f,$$

which means that $b<\phi_{1,\infty}\leqq 0$. The lemma is proved.

As said before, it follows from (h_3) that there are unique numbers α and β, $b<\beta<\alpha<0$, satisfying

$$\frac{k_M}{k_m}\int_0^a f=\int_0^\beta f \quad\text{and}\quad \frac{k_m}{k_M}\int_0^a f=\int_0^\alpha f.$$

Next we prove that $b<\beta\leqq\phi_{1,\infty}(1)\leqq\alpha<0$. To this end we need some additional lemmas.

LEMMA 2.8. *With notation set forth above and under hypotheses* (h_1), (h_2), *and* (h_3) *it holds that*

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\int_0^1 rk(r)\phi_{1,r}^2(r,\lambda)\,dr=0.$$

Proof. Since $\phi_1(r,\lambda)$ is an equilibrium solution to (2.1) it satisfies

$$\frac{1}{\lambda}\int_0^1 rk(r)\phi_{1,r}^2(r,\lambda)\,dr=\int_0^1 r\phi_1(r,\lambda)f[\phi_1(r,\lambda)]\,dr.$$

The second integral in the above inequality has a uniformly bounded integrand in $[0,1]$, and Lemma 2.6 assures us that $\phi_1(r,\lambda)\to a$ in measure, as $\lambda\to\infty$, for $0\leqq r\leqq 1$. Then an application of the Lebesgue Convergence Theorem yields the desired result.

LEMMA 2.9. *Under hypotheses* (h_1), (h_2), *and* (h_3) *it holds that any sequence* $\{\lambda_n\}_{n\in N}$, $\lambda_n\to\infty$ *as* $n\to\infty$, *has a subsequence* $\{\lambda_{n_k}\}_{k\in N}$ *such that*

$$\lim_{k\to\infty}\frac{1}{\lambda_{n_k}}\int_0^1\frac{[k(r)\phi_{1,r}(r,\lambda_{n_k})]^2}{r}\,dr=0.$$

Proof. To simplify notation we set

$$g_n(r)=\frac{rk(r)\phi_{1,r}^2(r,\lambda_n)}{\lambda_n}$$

for $0 \leqq r \leqq 1$. Then Lemma 2.8 says that

$$\lim_{n\to\infty} \int_0^1 g_n(r)\,dr = 0.$$

Therefore, all we have to prove is that

$$\lim_{n\to\infty} \int_0^1 g_n(r) \frac{k(r)}{r^2}\,dr = 0.$$

If we call

$$h_m(r) = \frac{k(r)}{r^2} \chi_{[1/m,1]}(r),$$

for $0 \leqq r \leqq 1$, $m \in N$, $m \geqq 1$, where $\chi_{[1/m,1]}$ stands for the characteristic function of $[1/m, 1]$, then $h_m(r) \to h(r)$, pointwise in $[0, 1]$, as $m \to \infty$, where $h(r) = k(r)/r^2$ for $r \in (0, 1]$ and $h(0) = 0$.

Let us call $\delta_{nm} = \int_0^1 h_m(r) g_n(r)\,dr$. Then

$$\begin{aligned} 0 < \delta_{nm} &= \int_0^1 g_n(r) \frac{k(r)}{r^2} \chi_{[1/m,1]}(r)\,dr \\ &\leqq \int_0^1 \frac{k^2(r)\phi_{1,r}^2(r, \lambda_n)}{\lambda_n r}\,dr = -\int_0^1 k(r) \frac{d}{dr} F[\phi_1(r, \lambda_n)]\,dr \\ &\leqq k_M \int_0^1 \left| \frac{d}{dr} F[\phi_1(r, \lambda_n)] \right| dr = k_M V_{F[\phi_1]}[0, 1] \\ &= k_M V_F[\phi_1(1, \lambda_n), \phi_1(0, \lambda_n)] \leqq k_M V_F[a, b] < \infty. \end{aligned}$$

In the first equality above we used an equality derived in the proof of Lemma 2.1. Also $V_f[a, b]$ stands for the total variation of f over $[a, b]$, and $V_F[a, b] < \infty$ follows from the fact that $f \in C^2(R, R)$.

So the double real sequence satisfies $0 < \delta_{nm} < V_f[a, b] < \infty$, and hence it has a subsequence (labelled again δ_{nm}) that converges.

Moreover, for a fixed m_0, we have

$$\lim_{n\to\infty} \delta_{nm} = \lim_{n\to\infty} \int_0^1 h_{m_0}(r) g_n(r) \leqq k_M m_0^2 \lim_{n\to\infty} \int_0^1 g_n(r)\,dr = 0.$$

As a consequence the two limits can be interchanged, yielding $\lim_{n\to\infty} \lim_{m\to\infty} \delta_{nm} = \lim_{m\to\infty} \lim_{n\to\infty} \delta_{nm} = 0$. Next, using the monotone convergence theorem, we obtain

$$\begin{aligned} \lim_{n\to\infty} \int_0^1 h(r) g_n(r)\,dr &= \lim_{n\to\infty} \lim_{m\to\infty} \int_0^1 g_n(r) h_m(r)\,dr \\ &= \lim_{m\to\infty} \lim_{n\to\infty} \delta_{nm} = 0 \quad \text{by the above calculation.} \end{aligned}$$

The lemma is proved.

LEMMA 2.10. *Let $\{\lambda_n\}_{n \in N}$ be such that $\lambda_n \to \infty$ as $n \to \infty$ and $\phi_1(1, \lambda_n)$ converges. Then*

$$k_m \lim_{n\to\infty} \int_0^{\phi_1(0,\lambda_n)} f \leqq k_M \lim_{n\to\infty} \int_0^{\phi_1(1,\lambda_n)} f.$$

Proof. Since $0 < \phi_1(1, \lambda) < b$ for any $\lambda \geqq \bar{\lambda}_1$, a sequence as above always exists.

It follows from the equality (derived in the proof of Lemma 2.1) that

$$\int_0^1 \frac{[k(r)\phi_{1,r}(r\lambda_n)]^2}{r}\,dr + \int_0^1 k(r)\frac{d}{dr}F[\phi_1(r,\lambda_n)]\,dr = 0.$$

This, along with Lemma 2.9, implies that

$$\lim_{k\to\infty}\int_0^1 k(r)\frac{d}{dr}F[\phi_1(r,\lambda_{n_k})]\,dr = 0$$

for any subsequence $\{\lambda_{n_k}\}_{k\in N}$ of the above sequence. For simplicity, we drop the subscript k in the notation of the subsequence.

With the notation set forth before we have

$$\begin{aligned}0 &\leqq \lim_{n\to\infty}\int_{r_1(\lambda_n)}^1 k(r)f[\phi_1(r,\lambda_n)]\phi_{1,r}(r,\lambda_n)\,dr\\ &= -\lim_{n\to\infty}\int_0^{r_1(\lambda_n)} k(r)f[\phi_1(r,\lambda_n)]\phi_{1,r}(r,\lambda_n)\,dr.\end{aligned}$$

A careful examination of the signs of the above integrands over the respective intervals of integration yields

$$\begin{aligned}0 &\leqq \lim_{n\to\infty}\left\{-k_m\int_0^{r_1(\lambda_n)}\frac{d}{dr}F[\phi_1(r,\lambda_n)]\,dr\right\}\\ &\leqq \lim_{n\to\infty}\left\{k_M\int_{r(\lambda_n)}^1 \frac{d}{dr}F[\phi_1(r,\lambda_n)]\,dr\right\},\end{aligned}$$

that is, $\lim_{n\to\infty} k_m \int_0^{\phi_1(0,\lambda_n)} f \leqq \lim_{n\to\infty} k_M \int_0^{\phi_1(1,\lambda_n)} f$. While comparing this to Lemma 2.1, note that the roles of $\phi_1(0,\lambda_n)$ and $\phi_1(1,\lambda_n)$ have been inverted here.

LEMMA 2.11. *Under the hypotheses of Lemma* 2.9, *and with α and β as before, and assuming the existence of*

$$\phi_{1,\infty}(1) = \lim_{n\to\infty}\phi_1(1,\lambda_n), \quad \textit{then } b < \beta \leqq \phi_{1,\infty}(1) \leqq \alpha < 0.$$

Proof. From Lemma 2.9 and previous results we obtain

$$k_m\int_0^a f \leqq k_M\int_0^{\phi_{1,\infty}(1)}.$$

But $\int_0^\alpha f = (k_m/k_M)\int_0^a f$ and, therefore, $\int_0^\alpha f \leqq \int_0^{\phi_{1,\infty}(1)} f$, that is, $b < \phi_{1,\infty}(1) \leqq \alpha < 0$.

On the other hand, Lemma 2.1 and the assumption that

$$\frac{k_M}{k_m}\int_0^a f = \int_0^\beta f,$$

where $b < \beta < 0 < a$, yield

$$\int_0^{\phi_{1,\infty}(1)} f \leqq \frac{k_M}{k_m}\int_0^a f = \int_0^\beta f < \int_0^b f.$$

Therefore, $\phi_{1,\infty}(1) \geqq \beta$, and, summing up, we have $b < \beta \leqq \phi_{1,\infty}(1) \leqq \alpha < 0$.

Note that if we knew that $\lim_{\lambda\to\infty}\phi(1,\lambda)$ existed, then there would be no need to resort to subsequences in the proof of Lemma 2.9.

COROLLARY 2.12. *With the above notation we have* $b < \beta \leqq \lim_{n\to\infty} \inf \phi_1(1, \lambda_n) \leqq \lim_{n\to\infty} \sup \phi_1(1, \lambda_n) \leqq \alpha < 0$ *for any sequence* $\{\lambda_n\}_{n\in N}$ *such that* $\lambda_n \to \infty$ *as* $n \to \infty$.

It follows by noting that $\lim_{n\to\infty} \inf \phi_1(1, \lambda_n)$ *and* $\lim_{n\to\infty} \sup \phi_1(1, \lambda_n)$ *are, respectively, the smallest and largest cluster points of the sequence* $\phi_1(1, \lambda_n)$.

COROLLARY 2.13. *If the diffusion function is constant, that is,* $k(r) = \text{const.}$, *then* $\lim_{\lambda\to\infty} \phi(1, \lambda) = c$, *where c is the unique negative number so that* $\int_0^a f = \int_0^c f$ *and* $b < c < 0 < a$.

Proof. It follows once we note that in this case $\int_0^a f = \int_0^\alpha f = \int_0^\beta f$ with $b < \alpha < 0$ and $b < \beta < 0$, and, therefore, $\alpha = \beta$ in view of hypothesis (h_2).

3. The case $\phi_1^-(., \lambda) \in C_1^-$: interior spike. This case can be handled in essentially the same way we handled the previous one, except for the order in which some lemmas are applied. Also, as a consequence of the fact that now $r_1(\lambda)$, the unique zero of $\phi_1(r, \lambda)$ in $(0, 1)$ satisfies $r_1(\lambda) \to 0$ as $\lambda \to \infty$; modifications must be introduced in some lemmas. We just indicate these modifications and how the proofs can be accomplished based on the foregoing results.

To start with we remark that in this case Lemma 2.1 provides insufficient information about the behavior of $r_1(\lambda)$ as $\lambda \to \infty$, since it does not contradict hypothesis (h_3). However, that can be overcome by adapting Lemma 2.9 to suit this case. Recall that $(d/dr)\phi_1^-(r, \lambda) > 0$ in $(0, 1)$.

LEMMA 3.1. *Suppose that* $\{\lambda_n\}_{n\in N}$ *is such that* $r_1(\lambda_n) \to \bar{r}$, *as* $n \to \infty$, *where* $0 \leqq \bar{r} \leqq 1$ *and* $\lim_{n\to\infty} \phi_1^-(0, \lambda_n)$ *and* $\lim_{n\to\infty} \phi_1^-(1, \lambda_n)$ *exist.*

Then under hypotheses (h_1), (h_2), *and* (h_3),

$$k_m \cdot \lim_{n\to\infty} \int_0^{\phi_1^-(0,\lambda_n)} f \leqq k_M \cdot \lim_{n\to\infty} \int_0^{\phi_1^-(1,\lambda_n)} f.$$

Proof. First, consider $0 < \bar{r} < 1$. Also take $\varepsilon > 0$, ε: small and set $r_\varepsilon^- = \bar{r} - (\varepsilon/2)$, $r_\varepsilon^+ = \bar{r} + (\varepsilon/2)$.

Then by constructing a supersolution to problem (2.1) in $[0, r_\varepsilon^-]$ and a subsolution to the same problem in $[r_\varepsilon^+, 1]$ we can follow mutatis mutandis the proof of Lemma 2.6 to conclude that $\phi_1^-(r, \lambda_n) \to b$ in measure, as $n \to \infty$, in $[0, \bar{r}]$ and $\phi_1^-(r, \lambda_n) \to a$, in measure, as $n \to \infty$, in $[\bar{r}, 1]$. In particular, for any $\varepsilon > 0$, $\phi_1^-(r, \lambda_n) \to b$, uniformly, as $n \to \infty$, in $[0, r_\varepsilon^-]$.

Also, since $\phi_1^-(r, \lambda_n)$ satisfies

$$\frac{1}{\lambda_n} \int_0^1 k(r)[\phi_{1,r}^-(r, \lambda_n)]^2 r \, dr = \int_0^1 \phi_1^-(r, \lambda_n) f[\phi_1^-(r, \lambda_n)] r \, dr,$$

we can use the Lebesgue Convergence Theorem to conclude that the limit as $n \to \infty$ of the left-hand side of the above equality is zero.

The same argument used in the proof of Lemma 2.9 yields

$$\lim_{n\to\infty} \frac{1}{\lambda_n} \int_0^1 \frac{[k(r)\phi_{1,r}^-(r, \lambda_n)]^2}{r} \, dr = 0 \quad \text{and, therefore,}$$

$$\lim_{n\to\infty} \left[\int_0^{r_1(\lambda_n)} k(r) \frac{d}{dr} F(\phi_1^-(r, \lambda_n)) \, dr + \int_{r_1(\lambda_n)}^1 k(r) \frac{d}{dr} F(\phi_1^-(r, \lambda_n)) \, dr \right] = 0.$$

Following the proof of Lemma 2.9, with natural adaptations, the desired result is obtained. The case $\bar{r} = 0$ as well as $\bar{r} = 1$ can be proved in a similar way.

Toward proving that $r_1(\lambda) \to 0$, as $\lambda \to \infty$, we suppose by contradiction that there is a sequence $\{\lambda_n\}_{n\in N}$ so that $r_1(\lambda_n) \to \bar{r}$, as $n \to \infty$, with $0 < \bar{r} \leqq 1$. Hence Lemma 3.1

can be evoked to conclude that $\lim_{n\to\infty} \phi_1^-(0, \lambda_n) = b$ and, therefore,

$$k_m \lim_{n\to\infty} \int_0^{\phi_1^-(0,\lambda_n)} f = k_m \int_0^b f \leqq k_M \lim_{n\to\infty} \int_0^{\phi_1^-(1,\lambda_n)} f \leqq k_M \int_0^a f,$$

which contradicts hypothesis (h_3).

Hence $r_1(\lambda) \to 0$ as $\lambda \to \infty$.

Also, following the proof of Lemma 2.6 we can show that $\phi_1^-(r, \lambda) \to a$, in measure, for $0 \leqq r \leqq 1$, as $\lambda \to \infty$. In particular $\phi_1^-(r, \lambda) \to a$, uniformly in $[\bar{r}, 1]$, as $\lambda \to \infty$ for any $\bar{r}$ such that $0 < \bar{r} < 1$.

Now Lemma 2.1, which also holds in this case, and our hypotheses assure us that if $\lim_{n\to\infty} \phi_1^-(0, \lambda_n) = \phi_{1,\infty}^-(0)$, then

$$\int_0^\alpha f = \frac{k_m}{k_M} \int_0^a f \leqq \int_0^{\phi_{1,\infty}^-(0)} f,$$

that is, $b < \phi_{1,\infty}^-(0) \leqq \alpha < 0$.

On the other hand, Lemma 3.1 yields

$$\int_0^{\phi_{1,\infty}(0)} f \leqq \frac{k_M}{k_m} \int_0^a f = \int_0^\beta f \quad \text{and so } b < \beta \leqq \phi_{1,\infty}^- \leqq \alpha < 0.$$

Likewise $b < \beta \leqq \lim_{n\to\infty} \inf \phi_1^-(0, \lambda_n) \leqq \lim_{n\to\infty} \sup \phi_1^-(0, \lambda_n) \leqq \alpha < 0$.

Note that since by the maximum principle $b < \phi_1^-(., \lambda) < a$ in $[0, 1]$, for any $\lambda > \tilde{\lambda}_1$, it follows that for any p, $1 \leqq p \leqq \infty$, $\lim_{\lambda\to\infty} \|\phi_1^-(., \lambda) - a\|_p = 0$, where as usual $\|.\|_p$ denotes the norm of the space $L^p[0, 1]$.

In view of the above conclusions it is easy to see that given any $\varepsilon > 0$, ε: small there is $\Lambda = \Lambda(\varepsilon)$ such that $b < \beta - \varepsilon < \phi_1^-(0, \lambda) < \alpha + \varepsilon < 0$ for $\lambda \geqq \Lambda$. See Fig. 2.

Remark 3.1. As for the general case $\phi_n^+(., \lambda) \in C_n^+[\phi_n^-(., \lambda) \in C_n^-]$, it still should hold, under hypotheses (h_1), (h_2), and (h_3), that the area determined by the negative

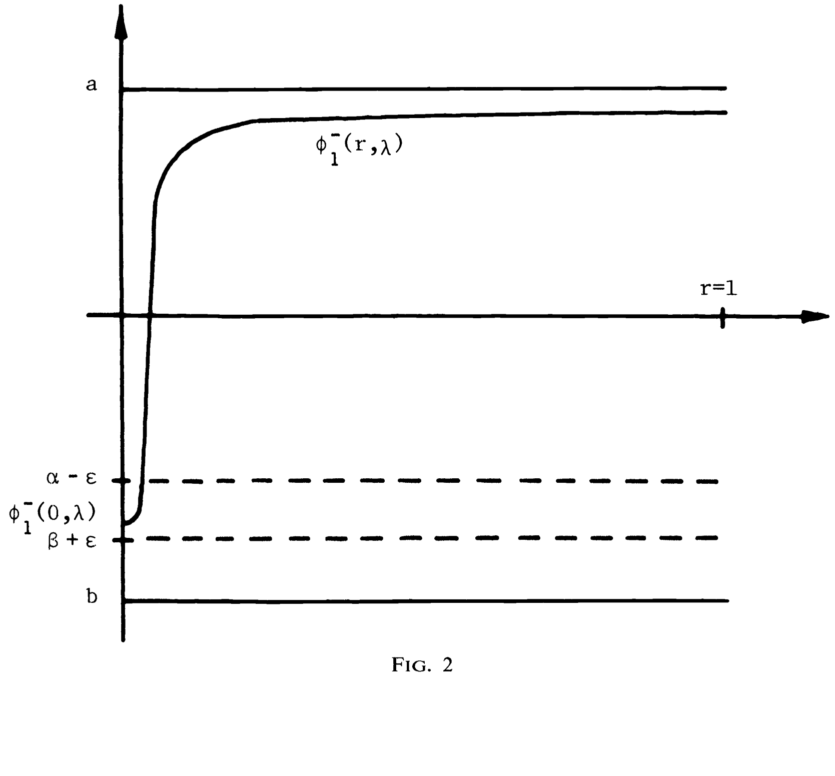


FIG. 2

bumps of $\phi_n^+(.,\lambda)[\phi_n^-(.,\lambda)]$ approaches zero as $\lambda\to\infty$. Let us point out that if $r_i(\lambda)$, $i=1,\dots,n$ stand for the zeros of $\phi_n^+(r,\lambda)$ $[\phi_n^-(r,\lambda)]$, $0<r<1$, where $0<r_1(\lambda)<\cdots< r_n(\lambda)<1$ and $\lambda>\tilde{\lambda}_n$, then there are unique $s_i(\lambda)$, $1\leqq i\leqq n-1$ such that $0<r_1(\lambda)< s_1(\lambda)<\cdots<r_{n-1}<s_{n-1}(\lambda)<r(\lambda)<1$, and

$$\frac{d}{dr}\phi_n^+(s_i)=0 \qquad \left[\frac{d}{dr}\phi_n^-(s_i)=0\right], \qquad 1\leqq i\leqq n-1.$$

Now it seems that the asymptotic behavior of $r_i(\lambda)$, $i=1,\dots,n$, for $n\geqq 2$ bears no resemblance to the case $n=1$. Actually the fact that the unique zero of $\phi_1^+(r,\lambda)$ $[\phi_1^-(r,\lambda)]$ in $(0,1)$ satisfies $r_1(\lambda)\to 1$ $[r_1(\lambda)\to 0]$ as $\lambda\to\infty$ might be a mere consequence of (h_3) forcing the area determined by negative part of $\phi_1^+(r,\lambda)$ $[\phi_1^-(r,\lambda)]$ in its graph shrink to zero as $\lambda\to\infty$ and should not be taken as a clue for determining the asymptotic behavior of $r_i(\lambda)$, $i=1,\dots,n$ for $n\geqq 2$. Instead, it seems that in the general case the focus should be shifted to the asymptotic behavior of $s_i(\lambda)$, $i=1,\dots,n-1$.

Remark 3.2. By reversing the inequality in hypothesis (h_3) it is possible to obtain the very same results of §§ 2 and 3 with the roles of the equilibria a and b switched.

Remark 3.3. Let us consider, instead of (1.1), the more general equation

$$\begin{gathered}\nabla\cdot[k(r)\nabla U(x)]+\lambda s(r)f(U(x))=0, \qquad x\in B,\\ \nabla U\cdot \hat{n}/_{\partial B}=0,\end{gathered} \tag{3.1}$$

where $r=\|x\|$ and $s(r)$ is a positive function in $C^2([0,1],R)$. The existence of two solutions, $U_1^+(r,\lambda)$ and $U_1^-(r,\lambda)$, of (3.1) having the same properties of the solutions $\phi_1^+(r,\lambda)$ and $\phi_1^-(r,\lambda)$ considered in this paper can be deduced from [1].

Set $s_m=\min_{0\leqq r\leqq 1} s(r)$, $s_M=\max_{0\leqq r\leqq 1} s(r)$, and let

$$s_m k_m\int_0^b f> s_M k_M\int_0^a f. \tag{h_3'}$$

Then under hypotheses (h_1), (h_2), and (h_3'), the techniques used in §§ 2 and 3 can be easily adapted to suit the above equation. Hence the same conclusions concerning the formation of boundary and interior spike formation can be drawn for the solutions $U_1^+(r,\lambda)$ and $U_1^-(r,\lambda)$, respectively.

REFERENCES

[1] A. S. DO NASCIMENTO, *Bifurcation and stability of radially symmetric equilibria of a parabolic equation with variable diffusion*, J. Differential Equations, 77 (1989), pp. 84–103.

[2] C.-S. LIN, W.-M. NI, AND I. TAKAGI, *Large amplitude stationary solutions to a chemotaxis system*, J. Differential Equations, 72 (1988), pp. 1–27.

[3] 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.

[4] ———, *Semilinear elliptic boundary value problems with small parameters*, Arch. Rational Mech. Anal., 52 (1973), pp. 205–232.

[5] F. A. HOWES, *Boundary-interior layer interactions in nonlinear singular perturbation theory*, Mem. Amer. Math. Soc., 15 (1978), pp. 1–108.

[6] P. H. RABINOWITZ, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487–513.

[7] D. G. DE FIGUEIREDO, *On the existence of multiple ordered solutions of nonlinear eigenvalue problems*, Nonlinear Anal. Theory Meth. Appl., 11 (1987), pp. 481–492.

[8] N. D. ALIKAKOS AND K. C. SHAING, *On the singular limit for a class of problems modelling phase transitions*, SIAM J. Math. Anal., 18 (1987), pp. 1453–1462.

[9] M. PROTTER AND H. WEIMBERGER, *Maximum Principles in Differential Equations*, Prentice Hall, Englewood Cliffs, NJ, 1967.

[10] Y. NISHIMURA, *Global structure of bifurcating solutions of some reaction-diffusion systems*, SIAM J. Math. Anal., 13 (1982), pp. 555-593.

[11] S. B. ANGEMENT, J. MALLET-PARRET, AND L. A. PELETIER, *Stable transition layers in a semilinear boundary value problem*, J. Differential Equations, 67 (1987), pp. 212-242.

[12] W.-M. NI AND I. TAKAGI, *On the existence and the shape of solutions to a semilinear Neumann problem*, 1989, preprint.

SIAM J. MATH. ANAL.
Vol. 24, No. 2, pp. 480–498, March 1993

SPECTRAL AND NONLINEAR EFFECTS IN CERTAIN ELLIPTIC SYSTEMS OF THREE VARIABLES*

LIGE LI† AND YAPING LIU†

Abstract. In this paper, the authors use the method of decomposing operator to prove the existence of positive solutions to certain elliptic biological interacting systems of three species in all possible patterns of interactions in terms of the combinations of competition, symbiosis, and predation under the homogeneous Robin–Dirichlet boundary conditions. The existence of positive solutions on large domains and the asymptotic stability of positive steady states for some cases are also proved. The main idea is that the existence of positive solutions and their properties can be characterized by the spectral properties of certain operators of Schrödinger type and by the equilibria of the system.

Key words. positive solutions, reaction-diffusion equations, steady states, principal eigenvalues of linear operators, equilibrium, spectral radius, bifurcation, stability

AMS(MOS) subject classifications. 35J60, 35P05, 47H10

1. Introduction.

In this paper we are concerned with the problem of positive solutions to nonlinear elliptic systems with three unknown functions. To better demonstrate the methods and the underlying ideas, we shall first investigate the following system:

$$(1)\qquad \begin{aligned} -\Delta u &= uf(u,v,w), \\ -\Delta v &= vg(u,v) && \text{in } \Omega, \\ -\Delta w &= wh(u,w), \\ B_1u = B_2v &= B_3w = 0 && \text{on } \partial\Omega, \end{aligned}$$

where Ω is a bounded domain in $\mathbf{R}^n$ with smooth boundary $\partial\Omega$ and $B_iu = a_i(x)u + b_i(x)\partial u/\partial\mathbf{n}$ are the Robin–Dirichlet boundary conditions with $a_i, b_i \in C(\mathbf{R}^n)$, $0 \not\equiv a_i(x) \geq 0, b_i(x) \geq 0$, and $a_i(x) + b_i(x) > 0$ on $\partial\Omega$ for $i = 1, 2, 3$. Moreover, f, g, h are C^1 monotone functions. This is the steady state of the reaction-diffusion system with initial data $\tilde{u}$, $\tilde{v}$, $\tilde{w}$:

$$(2)\qquad \begin{aligned} u_t - \Delta u &= uf(u,v,w), \\ v_t - \Delta v &= vg(u,v) && \text{in } (0,T)\times\Omega, \\ w_t - \Delta w &= wh(u,w), \\ B_1u = B_2v &= B_3w = 0 && \text{on } [0,T)\times\partial\Omega, \\ u(0,x) &= \tilde{u}(x), \\ v(0,x) &= \tilde{v}(x) && \text{in } \Omega, \\ w(0,x) &= \tilde{w}(x). \end{aligned}$$

Here u, v, w may represent the densities of interacting populations in problems arising from ecology, microbiology, immunology, etc. The functions f, g, h serve as the relative growth rates of these populations. We say that two species are in cooperation if each of their relative growth functions is increasing in the other; and that they are in competition if these functions are decreasing in the other opposer. In case of predation, one of the functions involved will be increasing in the prey while the other decreasing in the predator. For example, if w is a predator with u as its prey, then $f_w < 0$ and $h_u > 0$. See [40, Chap. 14] for details.

* Received by the editors March 18, 1991; accepted for publication (in revised form) July 9, 1992.

† Department of Mathematics, Kansas State University, Manhattan, Kansas 66506.

In §4, we shall extend the results on system (1) to the following more general elliptic system:

$$\begin{aligned} -\Delta u &= uf(u,v,w), \\ -\Delta v &= vg(u,v,w) &&\text{in } \Omega, \\ -\Delta w &= wh(u,w), \\ B_1 u = B_2 v &= B_3 w = 0 &&\text{on } \partial\Omega. \end{aligned} \tag{3}$$

The problem of positive solutions for single and two interacting species with diffusion can be traced back to the late 70s and early 80s. The pioneering investigations along this avenue include [2], [5], [9], [10], [11], [15], [16], [23], [24], [34], [36], [38], [39], [40], and many others. Since then a considerable amount of interest and intensive studies have been devoted to the systems of two interacting species, which may be termed as 2 × 2 systems in short. See [3], [4], [6], [7], [12], [14], [20], [21], [26], [29], [30], [35], [41], etc. In [16], [26], [37], [42] a good account of biological background was provided.

An important early discovery on the problem of positive coexistence of 2 × 2 systems is the following: the instability of the marginal densities, i.e., the individual species with the other species absent, implies the positive coexistence of both species provided that the interacting species are a priori bounded. See, for example, [2], [4], [9], [34], [37] for various predation and competition models. Later it was found that the stable marginal densities can also yield positive coexistence in competing interactions. See, for instance, [14], [30]. These two facts together tell us that it is the signs of the principal eigenvalues of the linearization of the PDE system at the related marginal densities that determine the positive coexistence when considering predation and competition. This includes the above instability principle. In case of symbiosis, the positive equilibria play an essential role. More specifically, for a system of two cooperative species, if the positive marginal densities exist and the system has a positive equilibrium, then it has a positive solution [21], [29]. We thus conclude that the positive coexistence for a 2 × 2 interacting system is determined by the spectral properties and the distribution of the equilibria of this system.

The problems arising from experimental and natural sciences, engineering, and technology often involve more than two components and challenge us to deal with 3 × 3 interacting systems. Recent research in neurobiochemistry has shown that the mechanism of the process of pain inducing and suppressing of human beings is closely related to the interactions between certain biochemical substances distributing along the neuropath nets. For example, it is known that two neurotransmitters, acetylcholine and triethylcholine, compete for chemical receptors. The first one can induce the pain while the second one serves as a suppressor. These two are both predated through deprivation of receptors by a third neuropeptide of a choline with a double group of

$$N^+\!\begin{matrix} /CH_3 \\ \backslash CH_3 \end{matrix}$$

in its structural formula (see [43]). The positive coexistence of such three species will affect the physiological feeling in an individual in response to a pain stimulation in terms of sensitivity and of tolerance. This problem can be modeled by system (1) or (3) with a suitable choice of the terms f, g, h.

As we have pointed out, the problem of the existence of positive solutions to 2 × 2 reaction-diffusion systems has been well investigated. However, the problem of three variables is far from being solved. To our knowledge, only some particular cases have been studied. A two predator, one prey model in a particular form of system (1) was considered in [42]. In [45], a special case of system (3) in a predator-prey-mutualist model was discussed. In [25] certain properties of positive solutions for a model of three species prey-predator system were investigated using the monotone scheme of upper-lower solutions. A general formulation of the mathematical problem of reaction-diffusion systems can be found in [16], [40]. In [31] the following system was studied:

$$\begin{aligned}\frac{\partial u_1}{\partial t} &= d_1\Delta u_1 + \alpha_1 u_1\left(1-\frac{u_1}{K_1}\right) - \frac{\alpha_1 u_1 u_2}{1+\beta u_3},\\ \frac{\partial u_2}{\partial t} &= d_2\Delta u_2 + \alpha_2 u_2\left(1-\frac{u_2}{K_2}\right) - \gamma u_1 u_2,\\ \frac{\partial u_3}{\partial t} &= d_3\Delta u_3 + \alpha_3 u_3\left(1-\frac{u_3}{1+u_1}\right).\end{aligned} \tag{4}$$

Here $d_i, \alpha_i, K_i, \beta, \gamma$ are positive constants. This is a model of two competitors with one cooperator. In [31] the authors applied the technique of upper-lower solutions to give certain conditions for the existence of positive solutions to system (4) and to its steady state under the homogeneous Dirichlet and Neumann boundary condition. In [22] the following system (one predator u, two preys v, w) was investigated:

$$\begin{aligned}-\Delta u &= u(a-u-cv-dw),\\ -\Delta v &= v(e+fu-v+gw) \qquad \text{in } \Omega,\\ -\Delta w &= w(\alpha-\beta u-\gamma v-w),\\ u &= v = w = 0 \qquad \text{on } \partial\Omega.\end{aligned} \tag{5}$$

Taking a or e as a bifurcation parameter, the author applied the fixed point index theory on positive cones to give a set of conditions under which the system (5) has positive solutions. In [17] the theory of fixed point index was also applied to a 3×3 predator-prey chain model.

The goal of this paper is to investigate the existence of positive solutions to 3×3 elliptic systems. As we mentioned in the foregoing paragraph about the 2×2 systems, the spectral properties of the linearization at marginal densities determine the positive coexistence in case of predations and competitions while it is dominated by the equilibria of the system in symbiotic interactions. We shall prove in this paper that the above principle for 2×2 systems carries over to 3×3 systems. More precisely, it may be formulated as follows. If the system is purely of symbiosis, then we have the positive coexistence provided that the system has a positive equilibrium (C_1, C_2, C_3), i.e., $f(C_1, C_2, C_3) = g(C_1, C_2) = h(C_1, C_3) = 0$ in the case of system (1) and $f(C_1, C_2, C_3) = g(C_1, C_2, C_3) = h(C_1, C_3) = 0$ in system (3). If the system involves only predation or competition or both, then the positive coexistence is characterized by the signs of the principal eigenvalues of certain differential operators that are determined by the marginal densities. That is, the spectral property of the system plays the primary role. If the system involves symbiosis together with predation and (or) competition, then the conditions for positive coexistence of three species will be a cocktail of the property of the equilibria and the spectral property of the system in question.

The results we are going to present in the following cover all possible interactions involving three species defined by the system (1) and (3), which include some of the earlier results as special cases. Moreover, the functions f, g, h are not specifically given and could be nonlinear. In this sense this paper gives a somewhat more general and systematic treatment of the problem on positive coexistence to certain general 3×3 elliptic systems. The methods employed in this paper are different from those used in the above mentioned papers. Our major approach is based on the method of decomposing operator, which is motivated by the idea used in [3]. In §2 we first improve this method by showing that the decomposing operators are C^1 mappings and their derivatives are either positive or negative operators. Then we combine this with the fixed point theory in order intervals due to [1] and with the bifurcation analysis to develop the techniques for the proofs.

Part of the results in §3 under the special case of Dirichlet boundary condition were presented in the AMC(1990) meeting and its proceeding.

2. Preliminaries. In this section we give several lemmas that are important in the sequel. In what follows, $\lambda_1(L)$ will denote the first eigenvalue of a suitable linear operator L and $r(L)$ the spectral radius of L. First we define the class $\mathcal{F} \subset C(\overline{\Omega} \times \mathbf{R})$ by the following.

DEFINITION 1. $p \in \mathcal{F}$ if and only if $p \in C(\overline{\Omega} \times \mathbf{R})$, p is C^1 in second component, $-L < p_u(x,u) < 0$ in $\Omega \times \mathbf{R}$ for some $L > 0$, and

$$\overline{\lim_{u \to +\infty}} p(x,u) < \lambda_1 \text{ uniformly for } x \in \Omega. \tag{6}$$

Here $\lambda_1 = \lambda_1(-\Delta)$ is the first eigenvalue of the equation

$$\begin{aligned} -\Delta u &= \lambda u \quad \text{in } \Omega, \\ Bu &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{7}$$

where $Bu = a(x)u + b(x)\partial u/\partial \mathbf{n}$ is the Robin boundary operator with $a, b \in C(\mathbf{R}^n), 0 \not\equiv a(x) \geq 0, b(x) \geq 0$, and $a(x) + b(x) > 0$ on $\partial\Omega$. The above inequality (6) will be satisfied if $p(x,C) < \lambda_1$ in Ω for some constant $C > 0$.

LEMMA 1. *Let $c(x) \in L^\infty(\Omega)$ and P be a positive constant such that $P + c(x) > 0$ for almost every $x \in \Omega$. Then*

(1) $\lambda_1(\Delta + c) > 0 \Leftrightarrow r[(-\Delta + P)^{-1}(P + c)] > 1$;
(2) $\lambda_1(\Delta + c) < 0 \Leftrightarrow r[(-\Delta + P)^{-1}(P + c)] < 1$;
(3) $\lambda_1(\Delta + c) = 0 \Leftrightarrow r[(-\Delta + P)^{-1}(P + c)] = 1$.

Proof. See Lemma 2 [29]. □

Let $\boldsymbol{K} = C(\overline{\Omega})^+$ be the positive cone of the ordered Banach space $C(\overline{\Omega})$. For $u_1, u_2 \in C(\overline{\Omega})$, define the order interval $[[u_1, u_2]] := \{u \in C(\overline{\Omega}) : u_1 \leq u \leq u_2 \text{ in } \Omega\}$. Let e be the unique solution of

$$\begin{aligned} -\Delta e &= 1 \quad \text{in } \Omega, \\ Be &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{8}$$

Then define the ordered Banach space $C_e(\overline{\Omega})$ by $C_e(\overline{\Omega}) = \bigcup_{\lambda \in \mathbf{R}^+} \lambda[[-e, e]] = \bigcup_{\lambda \in \mathbf{R}^+} [[-\lambda e, \lambda e]]$ with norm $\|u\|_e = \inf\{\lambda > 0 : -\lambda e \leq u \leq \lambda e\}$, the Minkowski functional. Let $\boldsymbol{K}_e := C_e(\overline{\Omega})^+$. (See [1] for details.)

LEMMA 2. *Let $p \in \mathcal{F}$. Consider*

$$\begin{aligned} -\Delta u &= up(x,u) \quad \textit{in } \Omega, \\ Bu &= 0 \quad \textit{on } \partial\Omega. \end{aligned} \tag{9}$$

(i) *If $\lambda_1(\Delta + p(x,0)) > 0$, then* (9) *has a unique positive solution.*

(ii) *If $\lambda_1(\Delta + p(x,0)) \le 0$, then $u \equiv 0$ is the only nonnegative solution of* (9).

Proof. (i) Let $\lambda_1(\Delta + p(x,0)) > 0$. Define operator $A : C(\overline{\Omega}) \to C(\overline{\Omega})$ by $Au = (-\Delta+P)^{-1}(up(x,u)+Pu)$, where $(-\Delta+P)^{-1}$ is under the boundary condition $Bu = 0$. $P > 0$ is to be chosen. Since $A'(0) = (-\Delta + P)^{-1}[p(x,0) + P]$ and $\lambda_1(\Delta+p(x,0)) > 0$, we have $r(A'(0)) > 1$ by Lemma 1. $A'(0)$ is strongly positive from $C(\overline{\Omega})$ to $C_e(\overline{\Omega})$ for $P > \max_{x\in\overline{\Omega}}\{-p(x,0)\}$. (See Theorem 4.2 in [1].) By Theorem 4.3 and Lemma 7.5 of [1], there exist $\epsilon \in (0,1]$ and $\underline{u} > 0$ such that $A(\tau\underline{u}) > \tau\underline{u}$ for all $\tau \in (0,\epsilon)$.

Let $\tilde{b}(x) = b(x)+\delta$ with $\delta > 0$, and let $\tilde{\lambda}_1 = \tilde{\lambda}_1(-\Delta)$ be the first eigenvalue under the boundary condition $\tilde{B}u = 0$, where $\tilde{B}u = au+\tilde{b}\partial u/\partial\mathbf{n}$. By the continuous dependence of eigenvalues on the boundary conditions (see, [13, Chap. 6] or use directly the variational formula for the principal eigenvalue), we chose δ small enough such that

$$\overline{\lim_{u\to+\infty}} p(x,u) < \tilde{\lambda}_1 \text{ uniformly for } x \in \Omega.$$

Let u be a positive eigenfunction corresponding to $\tilde{\lambda}_1$. Claim: $u > 0$ on $\overline{\Omega}$. $u > 0$ in Ω by the maximum principle. If $x_0 \in \partial\Omega$ is such that $u(x_0) = 0$, then u attains its nonpositive minimum at x_0. By the Hopf lemma (see [40, Chap. 8]), $\partial u/\partial\mathbf{n} < 0$ at x_0. On the other hand, $\partial u/\partial\mathbf{n} = -(a/\tilde{b})u = 0$ at x_0, a contradiction. Thus the claim is true and we can choose $\tilde{u} = uK$ with K large enough such that $\tilde{u} > \underline{u}$ and $p(x,\tilde{u}) < \tilde{\lambda}_1$ for all $x \in \Omega$. Therefore, $-\Delta\tilde{u} = \tilde{\lambda}_1\tilde{u} > \tilde{u}p(x,\tilde{u})$. Then $(-\Delta + P)\tilde{u} > \tilde{u}p(x,\tilde{u}) + P\tilde{u}$. We also have $B\tilde{u} \ge \tilde{B}\tilde{u} = 0$ on $\partial\Omega$ since $\tilde{b} > b \ge 0$ and $\partial\tilde{u}/\partial\mathbf{n} = -(a/\tilde{b})\tilde{u} \le 0$. Thus $\tilde{u} > (-\Delta + P)^{-1}[\tilde{u}p(x,\tilde{u}) + P\tilde{u}] = A\tilde{u}$ by the maximum principle. Let $C_0 = \max_{x\in\overline{\Omega}}\tilde{u}(x)$ and $P > 2\max\{\sup_{(x,u)\in\overline{\Omega}\times[0,C_0]}|p(x,u)|, C_0L\}$. Then A is an increasing, compact operator on the order interval $[[\tau\underline{u},\tilde{u}]]$ for a fixed $\tau \in (0,\epsilon)$. By Corollary 6.2 in [1] , A has a maximum fixed point $\overline{u} \in [[\tau\underline{u},\tilde{u}]]$. $\overline{u}$ is thus a strictly positive solution.

For the proof of the uniqueness see, for example, [17], [25].

(ii) Suppose u is a positive solution of (9). Then $(\Delta + p(x,u))u = 0$, and hence $\lambda_1(\Delta + p(x,u)) = 0$. Since $p(x,0) > p(x,u)$ in Ω, we have $\lambda_1(\Delta + p(x,0)) > \lambda_1(\Delta + p(x,u)) = 0$, a contradiction. □

Note 1. The above lemma generalizes Lemma 3 in [29] and the related result in [26] since we do not assume here the existence of a constant $C_0 > 0$ such that $p(x,C_0) \le 0$, for all $x \in \Omega$, and the boundary condition is also more general. When the applications to biology and chemical reaction are concerned, the assumption that $p(x,C_0) \le 0$ for some $C_0 > 0$ rules out many particularly important and useful models with mild change rates in the growth functions. $p(x,u) = cg(x)/(1+g(x)u)$ with $c > 0$ and $g(x) > 0$ is such an example.

In view of Lemma 2, for all $p \in \mathcal{F}$, let u_p be the unique positive solution of (9) if $\lambda_1(\Delta + p(x,0)) > 0$ and $u_p \equiv 0$ otherwise. The following result is motivated by the method in [3] §3 for a 2 × 2 predator-prey model.

LEMMA 3. *If the constant L in the definition of the class $\mathcal{F}$ is independent of functions p, then*

(i) *The mapping $p \mapsto u_p$ is continuous in sense of $\mathcal{F} \to C^{1,\alpha}(\Omega \times \mathbf{R})$ where $\alpha \in (0,1)$;*

(ii) *If $p_1 \ge p_2 \not\equiv p_1$, then either $u_{p_1} > u_{p_2}$ or $u_{p_1} \equiv u_{p_2} \equiv 0$.*

Proof. See Lemma 4 in [29] where it was proved for the Dirichlet boundary condition. But the arguments work also for the Robin boundary condition. See also [17]. □

LEMMA 4. *Let $c \in L^\infty(\Omega)$. If $\lambda_1(\Delta + c(x)) < 0$, where Δ is taken under the boundary condition $Bu = 0$ on $\partial\Omega$, then $(\Delta + c(x))^{-1}$ is a negative operator on $C(\overline{\Omega})$, i.e., it maps the positive cone $\boldsymbol{K}$ into $-\boldsymbol{K}$. More precisely, if $0 \not\equiv (\Delta + c(x))\phi \geq 0$, then $\phi < 0$.*

Proof. See Lemma 2.2 in [30], where it was proved for the Dirichlet boundary condition. But the method applys also to the Robin boundary condition as well because the maximum principle still applies. □

The following is a well-known result on a general version of the Hopf lemma.

LEMMA 5. *Assume $c(x) \in L^\infty(\Omega)$. Let $0 \not\equiv u \geq 0$ and $\Delta u + c(x)u \leq 0$ in Ω. If $u(x_0) = 0$ for a point $x_0 \in \partial\Omega$, then $\partial u(x_0)/\partial\mathbf{n} < 0$.*

Note 2. In this lemma no restriction on the sign of the function $c(x)$ is imposed due to the assumption $u \geq 0$. For a proof see [18, §1].

We impose the following hypotheses on the functions f, g, h:

(H1) $f \in C^1(\mathbf{R} \times \mathbf{R} \times \mathbf{R})$, $g, h \in C^1(\mathbf{R} \times \mathbf{R})$. $f(\cdot, v, w), g(u, \cdot), h(u, \cdot) \in \mathcal{F}$ for any fixed $u, v, w \in C(\overline{\Omega})^+$. $f_v, f_w, g_u, h_u \neq 0$ and all partial derivatives are uniformly bounded on $\mathbf{R} \times \mathbf{R} \times \mathbf{R}$ and on $\mathbf{R} \times \mathbf{R}$, respectively.

(H2) $\min\{f(0,0,0), g(0,0), h(0,0)\} > \lambda_1(-\Delta)$.

The hypothesis (**H1**) implies that the species u, v, w can survive by themselves in the absence of the other species, i.e., the equations

$$-\Delta u = uf(u,0,0) \quad \text{in } \Omega, B_1 u = 0 \quad \text{on } \partial\Omega,$$

$$-\Delta v = vg(0,v) \quad \text{in } \Omega, B_2 v = 0 \quad \text{on } \partial\Omega,$$

$$-\Delta w = wh(0,w) \quad \text{in } \Omega, B_3 w = 0 \quad \text{on } \partial\Omega,$$

have positive solutions u_0, v_0, w_0, respectively. This follows immediately from Lemma 2 since $\lambda_1(\Delta + f(0,0,0)) > 0$, $\lambda_1(\Delta + g(0,0)) > 0$, $\lambda_1(\Delta + h(0,0)) > 0$.

The functions u_0, v_0, w_0 will always denote the unique positive solutions of the above equations, respectively, in the sequel.

By hypothesis (**H1**), f_v, f_w, g_u, h_u do not change signs. So the pattern of interaction between each pair of two species is either predation, or symbiosis, or competition. We will use the following notation:

$$u \longrightarrow v \quad \text{if } u \text{ preys on } v,$$

$$u -\!\!\circ\!\!- v \quad \text{if } u, v \text{ cooperate},$$

$$u \longleftrightarrow v \quad \text{if } u, v \text{ compete}.$$

For example, $v \longleftarrow u \longrightarrow w$ represents a one predator(u) two prey(v, w) model, whereas $v -\!\!\circ\!\!- u \longleftrightarrow w$ means that u, v cooperate and u, w compete.

According to Lemma 2 we define the operators $S, T : C(\overline{\Omega}) \to C(\overline{\Omega})$ as follows. For $u \in C(\overline{\Omega})$, Su is the unique positive solution of the equation

$$\begin{aligned} -\Delta v &= vg(u,v) && \text{in } \Omega, \\ B_2 v &= 0 && \text{on } \partial\Omega, \end{aligned} \tag{10}$$

if $\lambda_1(\Delta + g(u,0)) > 0$ and $Su \equiv 0$ otherwise. Tu is determined similarly from the equation

$$
\begin{aligned}
-\Delta w &= wh(u,w) && \text{in } \Omega,\\
B_3 w &= 0 && \text{on } \partial\Omega.
\end{aligned}
\tag{11}
$$

Let $U_1 = \{u \in C(\overline{\Omega}) : Su > 0\}$, $U_2 = \{u \in C(\overline{\Omega}) : Tu > 0\}$. By Lemma 3, it is easy to see that the operators S and T are continuous operators and that if u, v cooperate or u is a prey of v, then S is a strictly increasing operator in the sense that $u_1 \not\equiv u_2 \geq u_1$ and $u_2 \in U_1$ implies $Su_2 > Su_1$, and that if u, v compete or u preys on v, then S is a strictly decreasing operator in the sense that $u_1 \not\equiv u_2 \geq u_1$ and $u_1 \in U_1$ implies $Su_2 < Su_1$. A similar conclusion is true for T with respect to U_2. Moreover, we have the following.

LEMMA 6. *The operators S, T defined above are C^1 mappings on U_1 and U_2, respectively. Moreover, the Fréchet derivative S' (or T') is either a positive operator or a negative operator.*

Proof. Define $F : C(\overline{\Omega}) \oplus C(\overline{\Omega}) \to C(\overline{\Omega})$ by $F(u,v) := -v + (-\Delta)^{-1}[vg(u,v)]$, where $(-\Delta)^{-1}$ is taken under the boundary condition $B_2 v = 0$. Let D_2 represent the derivative with respect to the second component. Then it can be verified that $D_2 F(u,v) = -I + (-\Delta)^{-1}[vg(u,v)]_v = -I + (-\Delta)^{-1}[g(u,v) + vg_v(u,v)]$, i.e., $D_2F(u,v)\phi = -\phi + (-\Delta)^{-1}[g(u,v)\phi + vg_v(u,v)\phi]$ for $\phi \in C(\overline{\Omega})$. Let $u^0 \in U_1$. We claim that $(D_2F)^{-1}$ exists and is bounded at (u^0, Su^0). It is equivalent to showing that 1 is not an eigenvalue of the operator $(-\Delta)^{-1}[vg(u^0,v)]_v$ at $v^0 = Su^0$. Suppose that there exists a $\phi \in C(\overline{\Omega})$, $\phi \neq 0$ such that $(-\Delta)^{-1}[v^0 g(u^0,v^0)]_v \phi = \phi$. Then $(\Delta + [v^0 g(u^0,v^0)]_v)\phi = 0 = 0 \cdot \phi$. This implies $0 \in \sigma(\Delta + [v^0 g(u^0,v^0)]_v)$. On the other hand, $-\Delta v^0 = v^0 g(u^0,v^0)$ implies $\lambda_1(\Delta + g(u^0,v^0)) = 0$. Since $g_v < 0$, it follows that $\lambda_1(\Delta + [v^0 g(u^0,v^0)]_v) < \lambda_1(\Delta + g(u^0,v^0)) = 0$. Thus $0 \notin \sigma(\Delta + [v^0 g(u^0,v^0)]_v)$. This contradiction shows that $D_2 F$ is nonsingular at (u^0, Su^0). Applying the implicit function theorem in Banach spaces (see Chapter 8 in [33]) to $F(u,v) = 0$, we see that there is a unique C^1 curve $v = \eta(u)$ in a neighborhood V of u^0 with $\eta(u^0) = Su^0$ and $F(u, \eta(u)) \equiv 0$. Note that $F(u,v) = 0$ if and only if v is a solution of the equation

$$
\begin{aligned}
-\Delta v &= vg(u,v) && \text{in } \Omega,\\
B_2 v &= 0 && \text{on } \partial\Omega.
\end{aligned}
\tag{12}
$$

Therefore, $Su = \eta(u)$ and S is a C^1 mapping from U_1 to $\boldsymbol{K}$. We, therefore, denote by $S'(u)$ the Fréchet derivative of S at u. Since $F(u,v) = -v + (-\Delta)^{-1}[vg(u,v)]$, we have $F_v = -I + (-\Delta)^{-1}(vg)_v$, $F_u = (-\Delta)^{-1} vg_u$. From $F(u, Su) = 0$ we conclude that $F_u + F_v S'(u) = 0$ by the chain rule. (See Chapter 6 in [33].) Therefore, $S'(u) = -F_v^{-1} F_u = -[-I + (-\Delta)^{-1}(vg)_v]^{-1}(-\Delta)^{-1} vg_u = -(\Delta + g + vg_v)^{-1} vg_u$, i.e., $S'(u) = -(\Delta + g + vg_v)^{-1}(vg_u)I$. Notice that $\lambda_1(\Delta + g) = 0$ because $(\Delta + g(u,v))v = 0$ for $v = Su > 0$. Therefore, $\lambda_1(\Delta + g + vg_v) < 0$ due to $g_v < 0$, and consequently $(\Delta + g + vg_v)^{-1}$ is a negative operator by Lemma 4. Since $v > 0$ and g_u has a definite sign, we conclude that $S'(u)$ is either a positive or a negative operator. The same conclusion can be drawn from the operator T. □

Note 3. If the functions g and h are C^2 functions, then the operators S and T are C^2 mappings [33].

With the operators S and T discussed above, we define and denote the operator $A : C(\overline{\Omega}) \to C(\overline{\Omega})$ by $Au = (-\Delta + P)^{-1}[uf(u, Su, Tu) + Pu]$, where $P > 0$ is a constant. Let $A_\theta = \theta A$ for $0 < \theta \leq 1$. Clearly, $\overline{u}$ is a fixed point of A in $\boldsymbol{K}$ if and only

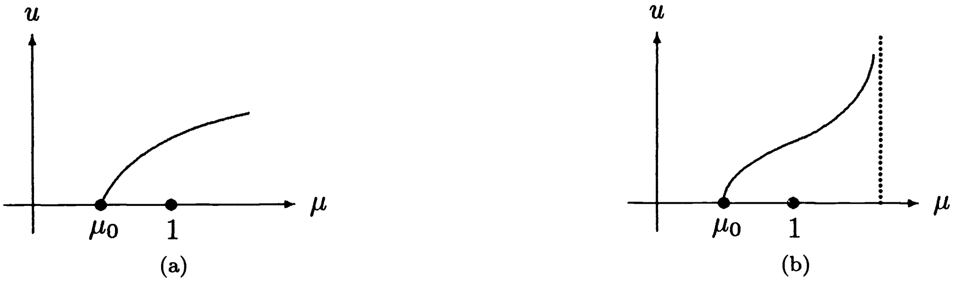


FIG. 1.

if $(\overline{u}, S\overline{u}, T\overline{u})$ is a nonnegative solution of (1). We are looking for the positive fixed point $\overline{u}$ of A with $S\overline{u} > 0$, $T\overline{u} > 0$.

LEMMA 7. *For a given $C > 0$, the operator A defined above is a positive, compact operator on $[[0, C]]$ for large P. If no predation is involved in system* (1), *then A is also increasing on $[[0, C]]$ for large P.*

Proof. Since $Au = (-\Delta + P)^{-1}[uf(u, Su, Tu) + Pu]$ and S, T are continuous, we see that A is compact. If $P > \sup_{u \in [[0,C]]} |f(u, Su, Tu)|$, then A is positive on $[[0, C]]$. Finally, if no predation is involved in system (1), then by Lemma 3 $f(u, Su, Tu)$ is increasing in u in the second and third components. Choose $P > \max_{u \in [[0,C]]} |D_1 f(u, Su, Tu)|$, where D_1 denotes the derivative with respect to the first component. Then A is an increasing operator on $[[0, C]]$. □

Let $\boldsymbol{X}$ be a Banach space, K be a compact linear strongly positive operator on $\boldsymbol{X}$ and F be a strictly positive continuous nonlinear operator on $\boldsymbol{X}$ with $F(0) = 0$. Consider the eigenvalue problem $u = \mu H u$ with $H = KF$, $(\mu, u) \in \mathbf{R} \times \boldsymbol{X}$. We have the following.

LEMMA 8. *Assume $H'(0)$ exists with $r(H'(0)) > 1$. If there exists no $\mu \in (0, 1]$ such that in its neighborhood the equation $u = \mu H u$ has solution u with $\|u\| \to \infty$, then H has a positive fixed point in $\boldsymbol{K}$.*

Proof. This is a simple consequence of the so-called nonlinear Krein–Rutman theorem (see, for example, Theorem 7.H in [44]), which says that for a strongly positive compact operator H, the $1/r(H'(0))$ is the only bifurcation point of μ for positive solutions $(\mu, u) > (0, 0)$, and that the bifurcation branch is unbounded. Note that for $\mu \leq 0$, the equation $u = \mu H u$ has no positive solution and the possibility of a bifurcation from ∞ in $(0, 1]$ is ruled out by our assumption. Therefore, the bifurcation diagram must be one of the cases in Fig. 1(a) and 1(b). Note that $\mu_0 = r(H'(0))^{-1} < 1$. The equation $u = \mu H u$ thus has a positive solution for each $\mu \in (\mu_0, 1]$. Let $\mu = 1$ to conclude the lemma. □

LEMMA 9. *The positive fixed points of A_θ have an a priori bound for $\theta \in (0, 1]$ if no symbiosis relation is involved in system* (1).

Proof. Let $A_\theta(u_\theta) = u_\theta$, $\theta \in (0, 1]$, $u_\theta \in \boldsymbol{K}$. Then

$$-\Delta u_\theta = \theta u_\theta f(u_\theta, Su_\theta, Tu_\theta) + P(\theta - 1) u_\theta. \tag{13}$$

Consider the following cases.

(1) Both v and w are predated by u. Then $f(u, Su, Tu) \leq f(u, C_2, C_3)$, where $C_2 = \max_{x \in \overline{\Omega}} v_0(x)$, $C_3 = \max_{x \in \overline{\Omega}} w_0(x)$. Also $f(0, C_2, C_3) \geq f(0, 0, 0) > \lambda_1(-\Delta)$.

(2) v is a prey for u while w is not. Then $f(u, Su, Tu) \leq f(u, C_2, 0)$ and $f(0, C_2, 0) \geq f(0, 0, 0) > \lambda_1(-\Delta)$.

(3) w is a prey for u while v is not. Then $f(u,Su,Tu) \le f(u,0,C_3)$ and $f(0,0,C_3) \ge f(0,0,0) > \lambda_1(-\Delta)$.

(4) Neither v nor w is a prey of u. Then $f(u,Su,Tu) \le f(u,0,0)$.

In any case, there exist $\overline{C}_2, \overline{C}_3 \ge 0$, independent of θ such that $f(u,Su,Tu) \le f(u,\overline{C}_2,\overline{C}_3)$ for any $u \in \boldsymbol{K}$ and $f(0,\overline{C}_2,\overline{C}_3) > \lambda_1(-\Delta)$. It thus follows from (13) that $-\Delta u_\theta \le \theta u_\theta f(u_\theta,\overline{C}_2,\overline{C}_3)$.

In case that $f(u,\overline{C}_2,\overline{C}_3) < 0$ as $u \ge C$ for some constant $C > 0$, the general maximum principle implies immediately that $u_\theta \le C$.

In case that $f(u,\overline{C}_2,\overline{C}_3) \ge 0$ for all $u \ge 0$, we find that u_θ is a lower solution of

$$(14)\qquad \begin{aligned} -\Delta u &= uf(u,\overline{C}_2,\overline{C}_3) && \text{in } \Omega,\\ B_1 u &= 0 && \text{on } \partial\Omega, \end{aligned}$$

because $0 < \theta \le 1$, $f \ge 0$, $u_\theta > 0$. Since we have already shown $f(0,\overline{C}_2,\overline{C}_3) > \lambda_1(-\Delta)$ and $\overline{\lim}_{u\to+\infty} f(u,\overline{C}_2,\overline{C}_3) < \lambda_1(-\Delta)$ by hypothesis (**H1**), as in the proof of Lemma 2, (14) has a unique positive solution $\overline{u}$, which is larger than or equal to any lower solution to (14). Thus $\overline{u} \ge u_\theta$. Note that $\overline{u}$ does not depend on θ. Therefore, $K = \max_{x\in\Omega} \overline{u}(x)$ is the a priori bound we need. □

PROPOSITION 1. *Assume that no symbiotic interaction is involved in system* (1). *Then A has a positive fixed point if $\lambda_1(\Delta + f(0,v_0,w_0)) > 0$.*

Proof. A is strongly positive from $C(\overline{\Omega})$ to $C_e(\overline{\Omega})$. (See Theorem 4.2 in [1] or Proposition 7.51 in [44]). $A'(0) = (-\Delta + P)^{-1}[f(0,v_0,w_0) + P]$ and $\lambda_1(\Delta + f(0,v_0,w_0)) > 0$ imply $r(A'(0)) > 1$. We apply Lemmas 9 and 8 to conclude this proposition. □

3. Results.

3.1. The existence theorem. In this section we give the conditions for the existence of positive solutions to the system (1) in all possible ten cases of interactions in three species in terms of the combinations of competition, symbiosis, and predation.

For $u_i, v_i, w_i \in C(\overline{\Omega})$ or $= +\infty, i = 1,2$, let $\ll (u_1,v_1,w_1),(u_2,v_2,w_2) \gg := \{(u,v,w) \in C(\overline{\Omega}) \oplus C(\overline{\Omega}) \oplus C(\overline{\Omega}) : (u_1,v_1,w_1) < (u,v,w) < (u_2,v_2,w_2) \text{ in } \Omega\}$.

Let u_0, v_0, w_0 denote, respectively, the unique positive density of species u, v, w while the other two species are absent. (See the paragraph following the statement of hypotheses (**H1**), (**H2**) in §2.)

THEOREM 1. *Let f,g,h satisfy* (**H1**) *and* (**H2**). *Then the following are true.*

Case 1. $v \longrightarrow u \longleftarrow w$. *System* (1) *has a positive solution if and only if* $\lambda_1(\Delta + f(0,v_0,w_0)) > 0$.

Case 2. $v -\!\!\circ\!\!- u -\!\!\circ\!\!- w$. *If system* (1) *has a positive equilibrium* (C_1,C_2,C_3), *then it has a positive solution in* $\ll (u_0,v_0,w_0),(C_1,C_2,C_3) \gg$.

Case 3. $v \longleftrightarrow u \longleftrightarrow w$. *Assume* $\lambda_1(\Delta + f(0,v_0,w_0)) > 0$, $\lambda_1(\Delta + g(u_0,0)) > 0$, $\lambda_1(\Delta + h(u_0,0)) > 0$; *then system* (1) *has a positive solution in*$\ll (0,0,0),(u_0,v_0,w_0) \gg$.

Note 4. *If one of* $\lambda_1(\Delta + g(u_0,0))$ *and* $\lambda_1(\Delta + h(u_0,0))$ *is* 0, *the conclusion is still true.*

Case 4. $v -\!\!\circ\!\!- u \longleftarrow w$. *Assume that* $\lambda_1(\Delta + f(0,v_0,w_0)) > 0$ *and the subsystem*

$$(15)\qquad \begin{aligned} -\Delta u &= uf(u,v,0),\\ -\Delta v &= vg(u,v), \end{aligned}$$

has a positive equilibrium (C_1,C_2), *i.e.*, $f(C_1,C_2,0) = 0, g(C_1,C_2) = 0$. *Then system* (1) *has a positive solution in* $\ll (0,v_0,w_0),(C_1,C_2,+\infty) \gg$.

Case 5. $v \longleftrightarrow u \longleftarrow w$. *Assume* $\lambda_1(\Delta + f(0, v_0, w_0)) > 0$, $\lambda_1(\Delta + g(u_0, 0)) \geq 0$. *Then system* (1) *has a positive solution in* $\ll (0, 0, w_0), (u_0, v_0, +\infty) \gg$.

Case 6. $v \mathrel{-\!\!\circ\!\!-} u \longleftrightarrow w$. *Assume that the subsystem*

$$\begin{aligned} -\Delta u &= uf(u, v, 0) \\ -\Delta v &= vg(u, v) \end{aligned} \tag{16}$$

has a positive equilibrium (C_1, C_2) *and that* $\lambda_1(\Delta + f(0, v_0, w_0)) > 0$, $\lambda_1(\Delta + h(C_1, 0)) \geq 0$. *Then system* (1) *has a positive solution in* $\ll (0, v_0, 0), (C_1, C_2, w_0) \gg$.

Note 5. *The above conclusion generalizes the existence result of Theorem* 2.7 *in* [31, §2], *where the condition* $\alpha_1(1 - K_2) > \lambda_0 d_1$ *implies* $\lambda_1(\Delta + f(0, 0, w_0)) > 0$ *in our term and* $\alpha_2 - \gamma K_1 > \lambda_0 d_2$ *implies* $\lambda_1(\Delta + h(C_1, 0)) > 0$.

Case 7. $v \longleftarrow u \longrightarrow w$. *Let* u_1 *be the unique positive solution of*

$$\begin{aligned} -\Delta u &= uf(u, v_0, w_0) && \text{in } \Omega, \\ B_1 u &= 0 && \text{on } \partial\Omega. \end{aligned} \tag{17}$$

(*Existence is guaranteed by Lemma* 2 *since* $\lambda_1(\Delta + f(0, v_0, w_0)) \geq \lambda_1(\Delta + f(0, 0, 0)) > 0$). *Assume that* $\lambda_1(\Delta + g(u_1, 0)) \geq 0$ *and* $\lambda_1(\Delta + h(u_1, 0)) \geq 0$. *Then system* (1) *has a positive solution in* $\ll (u_0, 0, 0), (u_1, v_0, w_0) \gg$.

Case 8. $v \mathrel{-\!\!\circ\!\!-} u \longrightarrow w$. *Let* $C = \max_{x \in \overline{\Omega}} w_0(x)$. *Assume*

$$\begin{aligned} -\Delta u &= uf(u, v, C), \\ -\Delta v &= vg(u, v) \end{aligned} \tag{18}$$

has a positive equilibrium (C_1, C_2) *and* $\lambda_1(\Delta + h(C_1, 0)) \geq 0$. *Then system* (1) *has a positive solution in* $\ll (u_0, v_0, 0), (C_1, C_2, w_0) \gg$.

Case 9. $v \longleftarrow u \longleftrightarrow w$. *Assume* $\lambda_1(\Delta + f(0, v_0, w_0)) > 0$ *and* $\lambda_1(\Delta + g(u_0, 0)) \geq 0$. *Let* u_1 *be the unique positive solution of*

$$\begin{aligned} -\Delta u &= uf(u, v_0, 0) && \text{in } \Omega, \\ B_1 u &= 0 && \text{on } \partial\Omega. \end{aligned} \tag{19}$$

(*Existence is guaranteed by Lemma* 2 *since* $\lambda_1(\Delta + f(0, v_0, 0)) \geq \lambda_1(\Delta + f(0, 0, 0)) > 0$.) *Suppose* $\lambda_1(\Delta + h(u_1, 0)) \geq 0$. *Then system* (1) *has a positive solution in* $\ll (0, 0, 0), (u_1, v_0, w_0) \gg$.

Case 10. $v \longleftarrow u \longleftarrow w$. *Assume* $\lambda_1(\Delta + f(0, v_0, w_0)) > 0$ *and* $\lambda_1(\Delta + g(u_0, 0)) \geq 0$. *Then system* (1) *has a positive solution in* $\ll (0, 0, w_0), (+\infty, v_0, +\infty) \gg$.

Note 6. *The condition* $\lambda_1(\Delta + g(u_0, 0)) \geq 0$ *can be weakened to* $\lambda_1(\Delta + g(u_1, 0)) \geq 0$, *where* u_1 *is the unique positive solution of*

$$\begin{aligned} -\Delta u &= uf(u, 0, w_0) && \text{in } \Omega, \\ B_1 u &= 0 && \text{on } \partial\Omega, \end{aligned} \tag{20}$$

if $\lambda_1(\Delta + f(0, 0, w_0)) > 0$ *and* $u_1 \equiv 0$ *otherwise. Thus, for example, if* $\lambda_1(\Delta + f(0, 0, w_0)) \leq 0$, *then* $\lambda_1(\Delta + f(0, v_0, w_0)) > 0$ *would imply the existence of a positive solution to system* (1).

Proof.

Case 1. In this case $f_v, f_w < 0$, $g_u, h_u > 0$. Let $Au := (-\Delta + P)^{-1} u[f(u, Su, Tu) + P]$. Assume $\lambda_1(\Delta + f(0, v_0, w_0)) > 0$. By Proposition 1, A has a fixed point $\overline{u} > 0$. $S\overline{u} > S0 = v_0 > 0$, $T\overline{u} > T0 = w_0 > 0$. $(\overline{u}, S\overline{u}, T\overline{u})$ is thus a positive solution of system (1).

Conversely, let $(\overline{u}, \overline{v}, \overline{w})$ be a positive solution. Then $\overline{v} = S\overline{u} > S0 = v_0$, $\overline{w} = T\overline{u} > T0 = w_0$, and thus $f(0, v_0, w_0) > f(0, \overline{v}, \overline{w})$. Therefore, $\lambda_1(\Delta + f(0, v_0, w_0)) > \lambda_1(\Delta + f(0, \overline{v}, \overline{w})) > 0$ since the equation

$$(21) \qquad \begin{array}{ll} -\Delta u = uf(u, \overline{v}, \overline{w}) & \text{in } \Omega, \\ B_1 u = 0 & \text{on } \partial\Omega \end{array}$$

has positive solution $\overline{u}$.

Case 2. In this case $f_v, f_w, g_u, h_u > 0$. $-\Delta u_0 = u_0 f(u_0, 0, 0) < u_0 f(u_0, Su_0, Tu_0)$ implies $u_0 < Au_0$. Equation

$$(22) \qquad \begin{array}{ll} -\Delta v = vg(C_1, v) & \text{in } \Omega, \\ B_2 v = 0 & \text{on } \partial\Omega \end{array}$$

has a unique positive solution $v_1 = SC_1$. The constant C_2 is an upper solution of (22), and the function ϵv_1 is a lower solution to (22) for $\epsilon \in (0, 1]$. When ϵ is small, we have $\epsilon v_1 < C_2$. There thus exists a positive solution $\tilde{v}$ of (22) in $[[\epsilon v_1, C_2]]$. By the uniqueness, $\tilde{v} = v_1$. This shows that $SC_1 \leq C_2$. Similarly, $TC_1 \leq C_3$. So $-\Delta C_1 = 0 = C_1 f(C_1, C_2, C_3) \geq C_1 f(C_1, SC_1, TC_1)$. Thus $C_1 \geq AC_1$. By a similar argument, we see that $u_0 \leq C_1$. By Corollary 6.2 in [1] and Lemma 7, A has a positive fixed point $\overline{u} \in [[u_0, C_1]]$. Therefore, $(\overline{u}, S\overline{u}, T\overline{u})$ is a positive solution of (1). Also, $C_1 \not\equiv \overline{u} \leq C_1$ implies $v_0 = S0 < S\overline{u} < SC_1 \leq C_2$ and $w_0 = T0 < T\overline{u} < TC_1 \leq C_3$ by the strict monotonicity of the operators S and T. Since $-\Delta u_0 = u_0 f(u_0, 0, 0)$, $-\Delta\overline{u} = \overline{u} f(\overline{u}, S\overline{u}, T\overline{u})$, $-\Delta C_1 = 0 = C_1 f(C_1, C_2, C_3)$, and $f(u, 0, 0) < f(u, S\overline{u}, T\overline{u}) < f(u, C_2, C_3)$, we conclude that $u_0 < \overline{u} < C_1$ according to Lemma 3.

Case 3. In this case $f_v, f_w, g_u, h_u < 0$. By Proposition 1, A has a positive fixed point $\overline{u}$. By Lemma 3, $\overline{u} \leq u_0$. Thus $v_0 = S0 > S\overline{u} \geq Su_0 > 0$ and $w_0 = T0 > T\overline{u} \geq Tu_0 > 0$. Then $\overline{u} < u_0$. Therefore, $(\overline{u}, S\overline{u}, T\overline{u})$ is a positive solution.

Case 4. In this case $f_v, g_u, h_u > 0$, $f_w < 0$. Since $A'(0) = (-\Delta + P)^{-1}[f(0, v_0, w_0) + P]$ and $\lambda_1(\Delta + f(0, v_0, w_0)) > 0$, we have $r(A'(0)) > 1$. It is easy to check that the restriction of A on $C_e(\overline{\Omega})$ is strongly positive. By the Krein–Rutman theorem, $r(A'(0))$ is an eigenvalue of $A'(0)$ with a positive eigenvector and $A'(0)$ has no other eigenvalues with positive eigenvectors. Let $A_\theta = \theta A$ and $R = C_1$. Suppose $A_\theta u = u$ with $u \in \partial B_R(0) \cap \boldsymbol{K}$ for some $\theta \in (0, 1]$. Then $-\Delta u = \theta uf(u, Su, Tu) + P(\theta - 1)u$. Since u, Su, and Tu are in $W^{2,p}(\Omega)$ for any $p > 0$ and f is C^1, we see from the above equality that $u \in W^{3,p}(\Omega)$ for any $p > 0$. By the Sobolev imbedding theorem, $u \in C^2(\overline{\Omega})$. Now $u \leq C_1$ implies $Su \leq C_2$. If u attains its maximum at $x_0 \in \Omega$, let $u(x_0) = \max_{x \in \overline{\Omega}} u(x) = C_1 > 0$. Since $Tu > T0 = w_0 > 0$, we have $0 \leq -\Delta u(x_0) \leq \theta u(x_0) f(u(x_0), Su(x_0), Tu(x_0)) < \theta u(x_0) f(u(x_0), C_2, w_0(x_0)) \leq \theta C_1 f(C_1, C_2, 0) = 0$, a contradiction. If u attains its maximum only on the boundary of Ω, let $u(x_0) = \max_{x \in \overline{\Omega}} u(x) = C_1 > 0$ for some $x_0 \in \partial\Omega$. Because $\partial\Omega$ is smooth, we can choose a ball $B \subset \Omega$ such that $\overline{B} \cap \partial\Omega = \{x_0\}$. Since $f(u(x_0), Su(x_0), Tu(x_0)) < f(C_1, C_2, 0) = 0$, B can be chosen so small such that $f(u(x), Su(x), Tu(x)) < 0$ on $\overline{B}$. We have $-\Delta u(x) \leq \theta u(x) f(u(x), Su(x), Tu(x)) < 0$ on $\overline{B}$. Applying the Hopf lemma (Lemma 5) to u in the domain B, we conclude that $\partial u(x_0)/\partial \mathbf{m} > 0$ for any outward pointing direction $\mathbf{m}$ with respect to ∂B. Therefore, $\partial u(x_0)/\partial \mathbf{n} > 0$ with respect to $\partial\Omega$. Then $B_1 u(x_0) = a_1(x_0) u(x_0) + b_1(x_0) \partial u(x_0)/\partial \mathbf{n} > 0$, contradicting the boundary condition $B_1 u = 0$. In any case it shows that A_θ has no fixed point on $\partial B_R(0) \cap \boldsymbol{K}$ for $0 < \theta \leq 1$. Therefore, $Au \neq \lambda u$ for all $\lambda \in [1, \infty)$ and for all $u \in \partial B_R(0) \cap \boldsymbol{K}$. By Theorem 13.2 in [1], The operator A has a positive fixed point $\overline{u} \in [[0, C_1]]$. Now $(\overline{u}, S\overline{u}, T\overline{u})$ is a positive solution of (1). $T\overline{u} > T0 = w_0$. $C_1 \not\equiv \overline{u} \leq C_1$ implies

$v_0 = S0 < S\overline{u} < SC_1 \leq C_2$. Thus it follows from Lemma 3 that $\overline{u} < C_1$, using a similar argument to the proof of Case 2.

Case 5. In this case $f_v, f_w, g_u < 0$, $h_u > 0$. By Proposition 1, the operator A has a positive fixed point $\overline{u}$. Since $T\overline{u} > T0 = w_0 > 0$, we have by Lemma 3, $\overline{u} < u_0$. Thus $v_0 = S0 > S\overline{u} > Su_0 \geq 0$. Therefore, $(\overline{u}, S\overline{u}, T\overline{u})$ is in $\ll (0,0,w_0), (u_0, v_0, +\infty) \gg$.

Case 6. In this case $f_v, g_u > 0$, $f_w, h_u < 0$. That $\lambda_1(\Delta + f(0, v_0, w_0)) > 0$ implies $r(A'(0)) > 1$. Similarly as in the proof of Lemma 2, we can show that there exists $\underline{u} > 0$ with $\underline{u} < C_1$ and $A\underline{u} > \underline{u}$. Since $SC_1 \leq C_2$ and $TC_1 \geq 0$, we have $-\Delta C_1 = 0 = f(C_1, C_2, 0) \geq f(C_1, SC_1, TC_1)$. This shows $C_1 \geq AC_1$. By Lemma 7 and Corollary 6.2 in [1], A has a positive fixed point $\overline{u} \in [[\underline{u}, C_1]]$. $\overline{u} \neq C_1$ implies $0 \leq TC_1 < T\overline{u} < T0 = w_0$ and $0 < v_0 = S0 < S\overline{u} < SC_1 \leq C_2$. It follows that $\overline{u} < C_1$ by the argument used in the proof of Case 2. Therefore, $(\overline{u}, S\overline{u}, T\overline{u}) \in \ll (0, v_0, 0), (C_1, C_2, w_0) \gg$.

Case 7. In this case, $f_v, f_w > 0$, $g_u, h_u < 0$. By Proposition 1, the operator A has a positive fixed point $\overline{u}$. Since in this case $S\overline{u} < S0 = v_0$, $T\overline{u} < T0 = w_0$, we have $f(0, S\overline{u}, T\overline{u}) < f(0, v_0, w_0)$. Lemma 3 implies $\overline{u} < u_1$. Thus $S\overline{u} > Su_1 \geq 0, T\overline{u} > Tu_1 \geq 0$. Hence $u_0 < \overline{u}$ and $(\overline{u}, S\overline{u}, T\overline{u})$ is a positive solution to system (1).

Case 8. In this case $f_v, f_w, g_u > 0$, $h_u < 0$. Similarly, as in Case 4, the operator A has a positive fixed point $\overline{u} \in [[0, C_1]]$ and $\overline{u} \neq C_1$. Therefore, $v_0 = S0 < S\overline{u} < SC_1 \leq C_2$ and $w_0 = T0 > T\overline{u} > TC_1 \geq 0$. Thus $u_0 < \overline{u} < C_1$. $(\overline{u}, S\overline{u}, T\overline{u})$ is a positive solution of system (1).

Case 9. In this case $f_v > 0$, $f_w, g_u, h_u < 0$. By Proposition 1, A has a positive fixed point $\overline{u}$. $S\overline{u} < S0 = v_0$ and $T\overline{u} \geq 0$ imply $f(0, S\overline{u}, T\overline{u}) < f(0, v_0, 0)$. Lemma 2 then asserts that $\overline{u} < u_1$. Then $0 \leq Tu_1 < T\overline{u} < T0 = w_0$. By Lemma 2, either $S\overline{u} > 0$ or $S\overline{u} \equiv 0$. Suppose $S\overline{u} \equiv 0$, then $\overline{u}$ is the positive solution of

$$(23) \qquad \begin{aligned} -\Delta u &= uf(u, 0, T\overline{u}) && \text{in } \Omega, \\ B_1 u &= 0 && \text{on } \partial\Omega. \end{aligned}$$

The inequality $f(0, 0, T\overline{u}) < f(0, 0, 0)$ implies $\overline{u} < u_0$ by Lemma 3. Hence $S\overline{u} > Su_0 \geq 0$, a contradiction. Therefore, $S\overline{u} > 0$. $(\overline{u}, S\overline{u}, T\overline{u})$ is a positive solution to system (1).

Case 10. In this case $f_v, h_u > 0$, $f_w, g_u < 0$. By Proposition 1, the operator A has a positive fixed point $\overline{u}$. $S\overline{u} < S0 = v_0$, $T\overline{u} > T0 = w_0 > 0$. Similarly, as in Case 9, we can show that $S\overline{u} > 0$. Consequently, $(\overline{u}, S\overline{u}, T\overline{u})$ is a positive solution of system (1) in $\ll (0, 0, w_0), (+\infty, v_0, +\infty) \gg$. □

What follows is a result on the existence of positive solutions over large domains. In addition to the hypothesis **(H1)**, we assume that there are constants $\overline{C}_i > 0$, $i = 2, 3$ such that $g(0, \overline{C}_2) = 0$, $h(0, \overline{C}_3) = 0$. The hypothesis **(H2)** will be replaced by the following weakened one.

(H2′) $f(0,0,0) > 0$, $g(0,0) > 0$, and $h(0,0) > 0$.

We say that a domain Ω is large if Ω contains a ball $B_R(0)$ with large radius R. Then we have the following.

PROPOSITION 2. *Under the above assumptions on f, g, h, system* (1) *in the case of* $(v \longrightarrow u \longleftarrow w)$ *has a positive solution on large domains if it has a positive equilibrium* (C_1, C_2, C_3).

Proof. We first have $C_2 > \overline{C}_2$ because $g(C_1, C_2) = g(0, \overline{C}_2) = 0$ and $g_u > 0$, $g_v < 0$. Similarly, $C_3 > \overline{C}_3$. So $f(0, \overline{C}_2, \overline{C}_3) > f(C_1, C_2, C_3) = 0$. We claim that when Ω is large, there exist unique positive solutions $v_0 > 0, w_0 > 0$, respectively, to the equations

$$\begin{aligned} &-\Delta v = vg(0,v) \quad \text{in } \Omega, B_2 v = 0 \quad \text{on } \partial\Omega, \\ &-\Delta w = wh(0,w) \quad \text{in } \Omega, B_3 w = 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Because $\lambda(-\Delta) \sim 1/(d(\Omega))^n$, where $d(\Omega)$ is the diameter and n the dimension of the domain Ω, the $d(\Omega)$ can be chosen so large that $\min\{\lambda_1(\Delta+g(0,0)), \lambda_1(\Delta+h(0,0))\} > 0$. Lemma 2 then asserts the existence of v_0, w_0. We have $v_0 < \overline{C}_2$, $w_0 < \overline{C}_3$ since $\overline{C}_2, \overline{C}_3$ are the a priori bounds for v_0, w_0, respectively, on every domain by the general maximum principle. Thus $f(0, v_0(x), w_0(x)) > f(0, \overline{C}_2, \overline{C}_3) > 0$ for all x. Therefore, $\lambda_1(\Delta + f(0, v_0, w_0)) > 0$ on large domains. Then apply Case 1 of Theorem 1 to complete the proof. □

Note 7. An interesting remark can be drawn by comparing this proposition with Case 1 in Theorem 1. That is, on an ordinary domain the spectral property of a two predator-one prey system determines the existence of positive solutions, while the equilibrium of the system dominates on a large domain. The solutions to an elliptic system over a large domain can be related to a singular perturbation problem, that is, a problem with a small diffusion rate. Under certain conditions the equilibrium might have a significant effect on the solutions. This idea had been adopted explicitly or implicitly by various authors. See, for example, [5], [8], [19], [28].

3.2. Bifurcations. Consider the following system

$$\begin{aligned} &-d\Delta u = uf(u,v,w), \\ &-\Delta v = vg(u,v) \qquad \text{in } \Omega, \\ &-\Delta w = wh(u,w), \\ &B_1 u = B_2 v = B_3 w = 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{24}$$

where $d > 0$ is a constant and the boundary operators $B_i u = a_i(x)u + b_i(x)\partial u/\partial \mathbf{n}$ are defined as in §1. We assume that f, g, h are C^2 functions, that (**H1**) is satisfied, and that $g(0,0), h(0,0) > \lambda_1(-\Delta)$. Let $C_{B_i}(\overline{\Omega})$ be the subspace of $C(\overline{\Omega})$, subject to the boundary condition $B_i u = 0$. Motivated by the method of [9] in two species of predator-prey relationship over an interval in $\mathbf{R}^1$, we give the uniqueness and stability of the positive solutions of (24) on various patterns of interactions.

THEOREM 2. *Under the above assumptions, if in* (24), $\lambda_1(d\Delta + f(0, v_0, w_0)) > 0$ *for some* $d > 0$, *then there is a neighborhood* $\mathcal{N}$ *of* $(0, v_0, w_0)$ *in* $C_{B_1}(\overline{\Omega}) \oplus C_{B_2}(\overline{\Omega}) \oplus C_{B_3}(\overline{\Omega})$ *such that there exists a unique positive solution* (u,v,w) *of system* (24) *in* $\mathcal{N}$ *for some range of d. This is true for all ten cases, and these solutions are stable in cases* 1, 7, 10 *in Theorem* 1.

Proof. Define $F(d,u) := d\Delta u + uf(u, Su, Tu)$, where the operators S and T are those defined in the paragraph right before Lemma 6 and $d > 0$ is the diffusion rate. It is easy to identify the linearization operator $F_u(d,0) = d\Delta + f(0, v_0, w_0)I$. The first eigenvalue of $F_u(d,0)$ depends smoothly on d for $d > 0$. Note that the spectral $\sigma(F_u(d,0))$ in $C_{B_1}(\overline{\Omega})$ coincides with that in $L^2(\overline{\Omega})$. Using the variational estimate of the selfadjoint operator $F_u(d,0)$ (under the boundary condition $B_1 u = 0$), we derive that $\lambda_1(F_u(d,0)) < 0$ for d large enough. Therefore, by the assumption of this theorem, there exists a $\overline{d} > 0$ for which $\lambda_1(F_u(\overline{d},0)) = 0$. Note that $F(d,0) \equiv 0$ for any given d. It is easy to see that $\overline{d}$ is a bifurcation point of d for the equation $F(d,u) = 0$. Indeed, let $\phi > 0$ be the corresponding eigenfunction of $F_u(\overline{d},0)\phi = 0$. Then an elementary calculation shows that $\operatorname{Ker}(F_u(\overline{d},0)) = \langle \phi \rangle$, the linear space generated by ϕ, and a function $\xi \in \operatorname{Range}(F_u(\overline{d},0))$ if and only if $\langle \phi, \xi \rangle = 0$, where

$\langle \cdot, \cdot \rangle$ denotes the inner product in $L^2(\Omega)$. Therefore, Codim(Range($F_u(\bar{d},0)$)) $= 1$. Finally, $F_{ud}(\bar{d},0) = \Delta$ and $F_{ud}\phi \notin$ Range($F_u(\bar{d},0)$) because $\langle \phi, \Delta\phi \rangle = -\int_\Omega |\nabla\phi|^2 dx + \int_{\partial\Omega} \phi\partial\phi/\partial\mathbf{n}ds < 0$ (otherwise $\phi \equiv 0$, which is contrary to $\phi > 0$). Thus the Crandall–Rabinowitz bifurcation theorem (see Theorem 13.5 in [40]) asserts the claim about $\bar{d}$. Therefore, we have a $\delta > 0$ and a C^1 bifurcating pair $d = d(s)$, $u = u(s) = s(\phi + \psi(s))$ with $d(0) = \bar{d}$, $\psi(0) = 0$, and $F(d(s), u(s)) = 0$ for $|s| < \delta$. We claim that $(u(s), Su(s), Tu(s))$ is a positive solution of system (24) for small $s > 0$. To this end we note first that $B_1\phi = a_1(x)\phi + b_1(x)\partial\phi/\partial\mathbf{n} = 0$ on $\partial\Omega$. For $x \in \partial\Omega$, if $b_1(x) = 0$, then $\phi(x) = 0$, and consequently $\partial\phi(x)/\partial\mathbf{n} < 0$ by Lemma 5. If $b_1(x) > 0$, then $\phi(x) > 0$. Otherwise $\phi(x) = 0$ implies $\partial\phi(x)/\partial\mathbf{n} = (a_1(x)/b_1(x))\phi(x) = 0$, which is contrary to Lemma 5. Let $F_1 = \{x \in \partial\Omega : b_1(x) = 0\}$. Then $\partial\phi(x)/\partial\mathbf{n} < 0$ on F_1. The continuity of b_1 implies that F_1 is compact. Thus there exists an open neighborhood U of F_1 and a positive number ϵ_1 such that $|\partial\phi(x)/\partial\mathbf{n}| > \epsilon_1$ on U. Let $F_2 = \partial\Omega \setminus U$. Then F_2 is compact and $\phi(x) > 0$ on F_2. Thus there exists an $\epsilon_2 > 0$ such that $\phi(x) > \epsilon_2$ on F_2. The a priori boundedness of the solution u implies that $u \in C^m(\bar{\Omega})$ for some $m > 1$ by the Sobolev imbedding theorem. Since $u(s) = s(\phi + \psi(s))$ and $\phi \in C^\infty$, it follows that $\psi(s) \in C^m(\bar{\Omega})$. Combining the fact that ψ is a C^1 mapping of s, we conclude that $\partial\psi(s)/\partial\mathbf{n}|_{\partial\Omega}$ is uniformly bounded for $|s| \leq \delta$. Putting all the above facts together with $\psi(s) = o(s)$, we can see that $u(s) = s(\phi + \psi(s)) > 0$ for small $s > 0$. Since $S0 = v_0 > 0$, $T0 = w_0 > 0$, and $u(s) \to 0$ as $s \to 0$, similar arguments can show that $Su(s) > 0$, $Tu(s) > 0$ for small $s > 0$.

The local uniqueness follows from that of the bifurcation solution.

In order to investigate the stability of the bifurcating solution, we first notice that the eigenfunction $\phi \notin$ Range($F_u(\bar{d},0)$) $= \{\xi : \langle \xi, \phi \rangle = 0\}$, and thus can apply the Rabinowitz theorem on the stability of bifurcating solutions (see, for example, Theorem 13.8 in [40]). In our case it claims that

$$\lim_{s\to 0} \frac{sd'(s)\lambda_1'(\bar{d})}{\eta(s)} = -1, \tag{25}$$

where the function $\lambda_1(d) = \lambda_1(F_u(d,0))$ is the principal eigenvalue of the operator $F_u(d,0)$, while $\eta(s)$ is the principal eigenvalue corresponding to the linearization of the bifurcating solution $u(s)$. By the variational estimate of the principal eigenvalue, it is clear that $\lambda_1'(d) < 0$. Let us estimate $d'(s)$. We have $-d(s)\Delta u(s) = u(s)f(u(s), Su(s), Tu(s))$. Namely, $-d(s)\Delta(s(\phi + \psi(s))) = s(\phi + \psi(s))f(s(\phi + \psi(s)), S(s(\phi+\psi(s))), T(s(\phi+\psi(s))))$. Dividing s on both sides and then taking the derivative with respect to s and evaluating at $s = 0$, it gives

$$\begin{aligned} &-d'(0)\Delta\phi - \bar{d}[\Delta\psi'(0)] \\ &= \psi'(0)f(0, v_0, w_0) + \phi[f_u(0, v_0, w_0)\phi + f_v(0, v_0, w_0)S'(0)\phi + f_w(0, v_0, w_0)T'(0)\phi]. \end{aligned}$$

The existence of the Fréchet derivative of the left-hand side is due to that of the right-hand side. Multiplying ϕ on both sides and then taking integral by part, we come up with the following equation:

$$-d'(0)\int_\Omega (\Delta\phi)\phi dx = \int_\Omega \phi^2[f_u\phi + f_vS'(0)\phi + f_wT'(0)\phi]dx. \tag{26}$$

Here we have used the cancellation

$$\langle \bar{d}\Delta\psi'(0) + \psi'(0)f(0, v_0, w_0), \phi \rangle = \langle \psi'(0), \bar{d}\Delta\phi + f(0, v_0, w_0)\phi \rangle = \langle \psi'(0), 0 \rangle = 0.$$

Note that by the proof of Lemma 6, $S'(u) = -(\Delta + g + vg_v)^{-1}(vg_u)I$, and $T'(u) = -(\Delta + h + wh_w)^{-1}(wh_u)I$, where $g = g(u(s), v(s))$, $h = h(u(s), w(s))$, $v(s) = Su(s)$, $w(s) = Tu(s)$. Notice also that $\lambda_1(\Delta + g) = \lambda_1(\Delta + h) = 0$. Therefore, $\lambda_1(\Delta + g + vg_v) < 0$ and $\lambda_1(\Delta + h + wh_w) < 0$ since $g_v < 0$, $h_w < 0$. First of all we consider the case of $v \longleftarrow u \longrightarrow w$ (Case 7 in Theorem 1). With the fact that $g_u < 0$ and $h_u < 0$ in mind, Lemma 4 then concludes that $S'(u)$ and $T'(u)$ are negative operators. Hence the right-hand side of (26) is negative because $f_u\phi < 0$, $f_v > 0$, $f_w > 0$, and $\phi > 0$. We then obtain immediately that $d'(0) < 0$. Therefore, $d'(s) < 0$ for small s. As a consequence, the function $\eta(s)$ in (25) must be negative for small positive s, and, therefore, the bifurcating solutions are stable. In case of $v \longrightarrow u \longleftarrow w$, $f_v < 0$, $f_w < 0$, $S'(u)$ and $T'(u)$ are positive operators; in case of $v \longleftarrow u \longleftarrow w$, $f_v > 0$, $f_w < 0$, $S'(u)$ is a negative operator while $T'(u)$ is a positive operator. Thus in each of these cases, the right-hand side of (26) is negative; therefore, $d'(s) < 0$ and the same conclusion follows. □

Note 8. To see that the above analysis on the equation $d\Delta u + uf(u, Su, Tu) = 0$ actually gives the bifurcation for system (24) at the point $(\bar{d}, (0, v_0, w_0))$, let us compute the spectrum of the linearization $\mathcal{L}$ of system (24) at $(0, v_0, w_0)$. It is easy to see that

(27)
$$\mathcal{L} = \begin{bmatrix} d\Delta + f(0, v_0, w_0) & 0 & 0 \\ v_0 g_u(0, v_0) & \Delta + g(0, v_0) + v_0 g_v(0, v_0) & 0 \\ w_0 h_u(0, w_0) & 0 & \Delta + h(0, w_0) + w_0 h_w(0, w_0) \end{bmatrix}.$$

Now $\mathcal{L} = \mathcal{L}_1 + B$, where $\mathcal{L}_1 = (d\Delta, \Delta, \Delta)^T$ and B is a bounded linear operator on $C_{B_1}(\bar{\Omega}) \oplus C_{B_2}(\bar{\Omega}) \oplus C_{B_3}(\bar{\Omega})$ given by (27) with the differential operators $d\Delta$, Δ, Δ being deleted. Because $\mathcal{L}$ has compact resolvent, the spectrum of $\mathcal{L}$ consists of only eigenvalues. Let $\mathcal{L}(\xi_1, \xi_2, \xi_3)^T = \lambda(\xi_1, \xi_2, \xi_3)^T$, where λ and $(\xi_1, \xi_2, \xi_3)^T$ is an eigenpair of the operator $\mathcal{L}$. It is easy to check that if $\xi_1 \neq 0$, then $\lambda \in \sigma(d\Delta + f(0, v_0, w_0))$. In case that $\xi_1 = 0$, letting $\xi_2 = 0$ or $\xi_3 = 0$ alternatively, we arrive at the relation $\sigma(\mathcal{L}) = \sigma(d\Delta + f(0, v_0, w_0)) \bigcup \sigma(\Delta + g(0, v_0) + v_0 g_v(0, v_0)) \bigcup \sigma(\Delta + h(0, w_0) + w_0 h_w(0, w_0))$. Note that the spectrums of the last two terms are in the subset $\{z : \text{Re}(z) \leq c < 0\}$ for some constant c. In order to have a bifurcation of system (24) we need the principal eigenvalue of $\mathcal{L}$ across zero. This can take place only for that of the operator $d\Delta + f(0, v_0, w_0)$, where d serves as a bifurcation parameter. This is equivalent to letting $\lambda_1(\bar{d}\Delta + f(0, v_0, w_0)) = 0$ as we did in the proof of Theorem 2.

Note 9. Theorem 2 gives the local results on the existence of positive solutions in terms of certain small range of the diffusion rate d, while Theorem 1 yields the global results, which are not subject to the diffusion rate.

4. Extensions. The methods developed in the previous sections apply also to system (3). The hypotheses **(H1)** and **(H2)** can be modified in the obvious way to accommodate the functions f, g, and h for this system. It is routine to verify that system (3) has a grand total of thirty-two possible nonequivalent types in combinations of the interactions between u, v, and w in terms of predation, competition and symbiosis. Of them eight are predations. First we need the following notation.

Let the functions u_0, v_0, and w_0 be as before. In view of the Lemma 2, let v^0 be the unique positive solution of the following equation

(28)
$$\begin{aligned} -\Delta v &= vg(0, v, w_0) && \text{in } \Omega, \\ B_2 v &= 0 && \text{on } \partial\Omega, \end{aligned}$$

if $\lambda_1(\Delta+g(0,0,w_0)) > 0$ and $v^0 \equiv 0$ otherwise. The functions $u_0 > 0$, $v_0 > 0$, $w_0 > 0$, and $v^0 \geq 0$, are totally determined by the shape of the domain Ω and the functions f, g, h in the system.

Notice that in system (3) the density v is determined by both u and w through the equation $-\Delta v = vg(u,v,w)$ and that v has no direct effect on the density w since $h_v \equiv 0$. By $w \longrightarrow v$ and $w \longleftarrow v$ we mean that $g_w < 0$ and $g_w > 0$, respectively.

From the last equation in system (3), we write $w = Tu$ as before, where $Tu > 0$ or $Tu \equiv 0$ depends on $\lambda_1(\Delta+h(u,0)) > 0$ or ≤ 0. Put $w = Tu$ in the second equation of (3); we can define the solution $v = Su$, where $Su > 0$ or $Su \equiv 0$ depends on whether $\lambda_1(\Delta + g(u,0,Tu)) > 0$ or ≤ 0. Plainly, $v^0 = S0$. Again, T is a strictly increasing or decreasing operator on the subset $U_2 = \{u \in C(\overline{\Omega}) : Tu > 0\}$, but the same conclusion may not be true for S in some cases. More precisely, S is strictly increasing on $U_1 = \{u \in C(\overline{\Omega}) : Su > 0\}$ if $g_u > 0$ and $g_w \cdot h_u > 0$, while S is strictly decreasing on U_1 if $g_u < 0$ and $g_w \cdot h_u < 0$. With the operators S and T in hand, we can define the operator A in the same way as before, $Au = (-\Delta + P)^{-1}[uf(u,Su,Tu) + Pu]$. The property of A depends on the particular pattern in system (3).

Lemma 9 is still true except for the case $v \longleftarrow u \longleftarrow w \longleftarrow v$. For this exceptional case we can impose certain mild conditions on the nonlinearities f, g, or h to keep Lemma 9 valid. For example, we assume that there exists a $C_0 > 0$ such that $h(u, C_0) \leq 0$ for any $u \geq 0$. Then for any positive solution w of

$$\begin{aligned} -\Delta w &= wh(u,w) && \text{in } \Omega, \\ B_3 w &= 0 && \text{on } \partial\Omega, \end{aligned} \tag{29}$$

$w \leq C_0$. We have the following.

LEMMA 9′. *The positive fixed points of A_θ have an a priori bound for $\theta \in (0,1]$ if no symbiosis relation is involved in system* (3).

PROPOSITION 1′. *Assume that no symbiotic interaction is involved in system* (3). *Then A has a positive fixed point if $\lambda_1(\Delta + f(0,v^0,w_0)) > 0$.* □

The proofs of Lemma 9′ and Proposition 1′ follow the same arguments as in that of Lemma 9 and Proposition 1, respectively.

The existence result is the following.

THEOREM 1′. *Let f, g, h satisfy the corresponding* (**H1**) *and* (**H2**). *Then the following are true.*

Case 1. $v \longrightarrow u \longleftarrow w \longleftarrow v$. *System* (3) *has a positive solution if and only if* $\lambda_1(\Delta + f(0,v^0,w_0)) > 0$.

Case 2. $v \multimap u \multimap w \longleftarrow v$. *If system* (3) *has a positive equilibrium* (C_1,C_2,C_3), *then it has a positive solution in* $\ll (u_0,v^0,w_0),(C_1,C_2,C_3) \gg$.

Case 3. $v \longleftrightarrow u \longleftrightarrow w \longrightarrow v$. *Assume* $\lambda_1(\Delta + f(0,v^0,w_0)) > 0$, $\lambda_1(\Delta + g(u_0,0,w_0)) \geq 0$ *and* $\lambda_1(\Delta + h(u_0,0)) \geq 0$. *Then system* (3) *has a positive solution in* $\ll (0,0,0),(u_0,v_0,w_0) \gg$.

Case 4. $v \multimap u \longleftarrow w \longleftarrow v$. *Assume that* $\lambda_1(\Delta + f(0,v^0,w_0)) > 0$ *and that the system*

$$\begin{aligned} -\Delta u &= uf(u,v,c), \\ -\Delta v &= vg(u,v,w), \\ -\Delta w &= wh(u,w) \end{aligned} \tag{30}$$

has a positive equilibrium (C_1,C_2,C_3), *i.e.,* $f(C_1,C_2,c) = g(C_1,C_2,C_3) = h(C_1,C_3) = 0$, *where* $c = \min_{x\in\overline{\Omega}} w_0(x)$. (*Note that if* $b_3(x) > 0$ *for all* $x \in \partial\Omega$ *in the boundary condition, then* $c > 0$.) *Then* (3) *has a positive solution in* $\ll (0,v^0,w_0),(C_1,C_2,C_3) \gg$.

Case 5. $v \,{-\!\!\circ\!\!-}\, u \longleftrightarrow w \longrightarrow v$. *Assume*

$$\begin{aligned} -\Delta u &= uf(u,v,0), \\ -\Delta v &= vg(u,v,0) \end{aligned} \tag{31}$$

has positive equilibrium (C_1, C_2) *and* $\lambda_1(\Delta + f(0,v^0,w_0)) > 0$, $\lambda_1(\Delta + g(0,0,w_0)) \geq 0$, $\lambda_1(\Delta + h(C_1,0)) \geq 0$. *Then* (3) *has a positive solution in* $\ll (0,v^0,0),(C_1,C_2,w_0) \gg$.

Proof.

Case 1. In this case $f_v, f_w < 0$, $g_u, g_w, h_u > 0$, and S is an increasing operator. Let $Au := (-\Delta + P)^{-1}u[f(u,Su,Tu) + P]$. Assume $\lambda_1(\Delta + f(0,v^0,w_0)) > 0$. By Proposition 1′, A has a fixed point $\overline{u} > 0$. $T\overline{u} > T0 = w_0 > 0$, $S\overline{u} > S0 = v^0 > v_0 > 0$. $(\overline{u}, S\overline{u}, T\overline{u})$ is thus a positive solution to system (3).

Conversely, let $(\overline{u},\overline{v},\overline{w})$ be a positive solution. Then $\overline{w} = T\overline{u} > T0 = w_0$, $\overline{v} = S\overline{u} > S0 = v^0$, $f(0,v^0,w_0) > f(0,\overline{v},\overline{w})$. Therefore, $\lambda_1(\Delta + f(0,v^0,w_0)) > \lambda_1(\Delta + f(0,\overline{v},\overline{w})) > 0$ since the equation

$$\begin{aligned} -\Delta u &= uf(u,\overline{v},\overline{w}) && \text{in } \Omega, \\ B_1 u &= 0 && \text{on } \partial\Omega \end{aligned} \tag{32}$$

has positive solution $\overline{u}$.

Case 2. In this case $f_v, f_w, g_u, g_w, h_u > 0$ and S is an increasing operator. It is similar to Case 2 of Theorem 1, showing that A has a positive fixed point $\overline{u} \in [[u_0, C_1]]$. Therefore, $(\overline{u}, S\overline{u}, T\overline{u})$ is a positive solution of (3). Also, $C_1 \not\equiv \overline{u} \leq C_1$ implies $w_0 = T0 < T\overline{u} < TC_1 \leq C_3$ and $v_0 < v^0 = S0 < S\overline{u} < SC_1 \leq C_2$. It then follows from Lemma 3 that $u_0 < \overline{u} < C_1$.

Case 3. In this case $f_v, f_w, g_u, g_w, h_u < 0$. The operator A has a positive fixed point $\overline{u}$ by Proposition 1′. $\overline{u} \leq u_0$ by Lemma 3. Therefore, $T\overline{u} < T0 = w_0$ and $S\overline{u} < v_0$. $\lambda_1(\Delta + g(\overline{u},0,T\overline{u})) > \lambda_1(\Delta + g(u_0,0,w_0)) \geq 0$. Thus $S\overline{u} > 0$. Then $\overline{u} < u_0$ and $T\overline{u} > Tu_0 \geq 0$. $(\overline{u}, S\overline{u}, T\overline{u})$ is a positive solution.

Case 4. In this case $f_v, g_u, g_w, h_u > 0$, $f_w < 0$, and the operators S and T are increasing. We argue similarly as in Case 4 of Theorem 1 that the operator A has a positive fixed point $\overline{u} \in [[0, C_1]]$. Now $(\overline{u}, S\overline{u}, T\overline{u})$ is a positive solution of (3). $T\overline{u} > T0 = w_0$. That $C_1 \not\equiv \overline{u} \leq C_1$ implies $T\overline{u} < TC_1 \leq C_3$. Thus $v_0 < v^0 = S0 < S\overline{u} < SC_1 \leq C_2$, and consequently we can show that $\overline{u} < C_1$ as we did in the proof of Case 4 in Theorem 1.

Case 5. In this case $f_v, g_u > 0$, $f_w, g_w, h_u < 0$ and S is an increasing operator. Similarly as in Case 6 of Theorem 1, we can show that A has a positive fixed point $\overline{u} \leq C_1$. $\overline{u} \not\equiv C_1$ implies $0 \leq TC_1 < T\overline{u} < T0 = w_0$ and $0 \leq v^0 = S0 < S\overline{u} < SC_1 \leq C_2$. Hence $\overline{u} < C_1$. Therefore $(\overline{u}, S\overline{u}, T\overline{u}) \in \ll (0,v^0,0),(C_1,C_2,w_0) \gg$. □

For the details on the statements and proofs of the other twenty-seven cases of system (3), and for Case $h = h(u,v,w)$ in the system (3), see [32].

The analogue of Theorem 2 is the following.

THEOREM 2′. *Consider the following system:*

$$\begin{aligned} -d\Delta u &= uf(u,v,w), \\ -\Delta v &= vg(u,v,w) && \text{in } \Omega, \\ -\Delta w &= wh(u,w), \\ B_1 u = B_2 v &= B_3 w = 0 && \text{on } \partial\Omega, \end{aligned} \tag{33}$$

where we impose similar assumptions as in Theorem 2. *If* $\lambda_1(d\Delta + f(0,v^0,w_0)) > 0$ *for some* $d > 0$ *in system* (33), *then there is a neighborhood* $\mathcal{N}$ *of* $(0,v^0,w_0)$ *in*

$C_{B_1}(\overline{\Omega}) \oplus C_{B_2}(\overline{\Omega}) \oplus C_{B_3}(\overline{\Omega})$ *such that there exists a unique positive solution* (u, v, w) *of system* (33) *in* $\mathcal{N}$ *for some range of* d. *This is true for all thirty-two patterns. These solutions are stable in the following four cases of predations:* $v \longrightarrow u \longleftarrow w \longleftarrow v$, $v \longleftarrow u \longrightarrow w \longleftarrow v$, $v \longrightarrow u \longrightarrow w \longrightarrow v$, *and* $v \longleftarrow u \longleftarrow w \longrightarrow v$.

The proof of this theorem is just a step-by-step imitation of that used for Theorem 2. □

Note 10. The four cases of predation stated in the last assertion of the above theorem are among the eight possible patterns of predation in system (33). In the other four cases the local bifurcating positive solutions are, in general, not stable.

Acknowledgment. The authors thank the referees for their very helpful commentary. They suggested the title for this paper and pointed out [34], [37], and many other references to us. The expository review in §1 is greatly under their influence.

REFERENCES

[1] H. Amann, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620–709.

[2] D. Aronson and H. Weinberger, *Nonlinear diffusion in population genetics*, Combustion and Nerve Propagation, Proc. Tulane Program in PDEs, Lecture Notes in Math. 446, Springer-Verlag, New York, 1975, pp. 5–49.

[3] 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.

[4] K. J. Brown, *Nontrivial solutions of predator-prey systems with small diffusions*, Nonlinear Anal. Theoret. Meth. Appl., 11 (1987), pp. 685–689.

[5] P. N. Brown, *Decay to uniform states in ecological interactions*, SIAM J. Appl. Math., 38 (1980), pp. 22–37.

[6] R. S. Cantrel and C. Cosner, *On the steady-state problem for the Lotka–Volterra competition model with diffusion*, Houston J. Math., 13 (1987), pp. 337–352.

[7] ———, *On the uniqueness and stability of positive solutions in the Lotka–Volterra competition model with diffusion*, Houston J. Math., 15 (1989), pp. 341–361.

[8] A. Castro, *Uniqueness of positive solutions for a sublinear Dirichlet problem*, Proc. Sympos. Pure Math., 45, Part 1, American Mathematical Society, Providence, RI, 1986, pp. 243–251.

[9] E. D. 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.

[10] E. D. Conway and J. Smoller, *A comparison technique for systems of reaction-diffusion equations*, Comm. Partial Differential Equations, 7 (1977), pp. 679–697.

[11] ———, *Diffusion and the classical ecological interaction*, in Nonlinear Diffusion, W. Fitzgibbon and H. Walker, eds., Pitman, London, 1977, pp. 53–69.

[12] 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.

[13] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. 1, Wiley-Interscience, New York, 1962.

[14] E. N. Dancer, *On positive solutions of some pairs of differential equations* II, J. Differential Equations, 60 (1985), pp. 236–258.

[15] O. Diekmann and N. Temme, *Nonlinear Diffusion Problems*, Math. Centrum Amsterdam, the Netherlands, 1976.

[16] P. C. Fife, *Mathematical aspects of reacting and diffusing systems*, Lecture Notes in Biomath. 28, Springer-Verlag, New York, Berlin, 1979.

[17] A. Ghoreishi and R. Logan, *Positive solutions of interacting models in heterogeneous environment of mixed boundary conditions*, Bull. Austral. Math. Soc., 44 (1991), pp. 79–94.

[18] B. Gidas, W. M. Ni, and L. Nirenberg, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.

[19] V. Hutson and K. Schmitt, *Permanence in dynamic systems*, Dynam. Report. Ser. Dynam. Syst. Appl., submitted.

[20] C. Keller and R. Lui, *Existence of steady-state solutions to predator-prey equations in a heterogeneous environment*, J. Math. Anal. Appl., 123 (1987), pp. 306–326.

[21] P. Korman and A. W. Leung, *On the existence and uniqueness of positive steady states in the Volterra–Lotka ecological models with diffusion*, Appl. Anal., 26 (1987), pp. 145–160.

[22] N. Lakoš, *Existence of steady-state solutions for an one predator-two prey system*, SIAM J. Math. Anal., 21 (1990), pp. 647–659.

[23] A. Lazer and J. McKenna, *On the number of solutions of a nonlinear Dirichlet problem*, J. Math. Anal. Appl., 84 (1981), pp. 282–294.

[24] A. Leung, *Monotone schemes for semilinear elliptic systems related to ecology*, Math. Meth. Appl. Sci., 4 (1982), pp. 272–285.

[25] ———, *A study of three species prey-predator reaction-diffusions by monotone schemes*, J. Math. Anal. Appl., 100 (1984), pp. 583–604.

[26] ———, *Systems of Nonlinear Partial Differential Equations: Applications to Biology and Engineering*, Klumer Academic Publishers, Norwell, MA, 1989.

[27] L. Li, *Coexistence theorms of steady-state for predator-prey interacting systems*, Trans. Amer. Math. Soc., 305 (1988), pp. 143–166.

[28] ———, *Global positive coexistence of a nonlinear elliptic biological interacting model*, Math. Biosci., 97 (1989), pp. 1–15.

[29] L. Li and A. Ghoreishi, *On positive solutions of general nonlinear elliptic symbiotic interacting systems*, Appl. Anal., 14 (1991), pp. 281–295.

[30] L. Li and R. Logan, *Positive solutions to general elliptic competition models*, J. Differential Integ. Eqs., 4 (1991), pp. 817–834.

[31] Z. Li, Z. Yang, and Q. Ye, *Some persistence and stability result for a competitor-competitor-mutualist model*, Universitatis Pekinensis, 24 (1988), pp. 1–12.

[32] Y. Liu, Ph.D. dissertation, Kansas State University, Manhattan, KS, 1993.

[33] L. A. Ljusternik and V. I. Sobolev, *Introduction to Functional Analysis*, Nauka, Moscow, 1965.

[34] D. Ludwig, D. G. Aronson, and H. F. Weinberger, *Spatial patterning of the spruce budworm*, J. Math. Biol., 8 (1979), pp. 217–258.

[35] P. J. McKenna and W. Walter, *On the dirichlet problem for elliptic systems*, Appl. Anal., 21 (1986), pp. 207–224.

[36] M. Mimura and J. D. Murray, *On a diffusion prey-predator model which exhibits patchiness*, J. Theoret. Biol., 75 (1979), pp. 249–262.

[37] S. W. Pacala and J. Roughgarden, *Spacial heterogeneity and interspecific competition*, Theoret. Population Biol., 21 (1982), pp. 92–113.

[38] C. V. Pao, *Coexistence and stability of a competition diffusion system in population dynamics*, J. Math. Anal. Appl., 83 (1981), pp. 54–76.

[39] ———, *On nonlinear reaction-diffusion systems*, J. Math. Anal. Appl., 87 (1982), pp. 165–198.

[40] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

[41] I. Stakgold, *Reaction-diffusion problems in chemical engineering*, Lecture Notes in Math. 1224, Springer-Verlag, New York, Berlin, 1986.

[42] P. Waltman, *Competition Models in Population Biology*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 45, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1984, p. 34.

[43] S. Xu, ed., *Neurobiochemistry*, Shanghai Medical University Press, Shanghai, 1989.

[44] E. Zeidler, *Nonlinear Functional Analysis and Its Applications* I, Springer-Verlag, New York, Berlin, 1986.

[45] S. Zheng, *A reaction-diffusion system of a predator-prey-mutualist model*, Math. Biosci., 78 (1986), pp. 217–245.

SIAM J. MATH. ANAL.
Vol. 24, No. 2, pp. 499-519, March 1993

014

ORTHONORMAL BASES OF COMPACTLY SUPPORTED WAVELETS II. VARIATIONS ON A THEME*

INGRID DAUBECHIES†

Abstract. Several variations are given on the construction of orthonormal bases of wavelets with compact support. They have, respectively, more symmetry, more regularity, or more vanishing moments for the scaling function than the examples constructed in Daubechies [*Comm. Pure Appl. Math.*, 41 (1988), pp. 909-996].

Key words. wavelets, orthonormal bases, regularity, symmetry

AMS(MOS) subject classifications. 26A16, 26A18, 26A27, 39B12

1. Introduction. This paper concerns the construction of orthonormal bases of wavelets, i.e., orthonormal bases $\{\psi_{jk};\, j, k \in \mathbb{Z}\}$ for $L^2(\mathbb{R})$, where

$$\psi_{jk}(x) = 2^{-j/2}\psi(2^{-j}x - k) \tag{1.1}$$

for some (very particular!) $\psi \in L^2(\mathbb{R})$. The functions (1.1) are *wavelets* because they are all generated from one single function by dilations and translations. Note that wavelets need not be orthogonal or even linearly independent. In fact, the "first" wavelets were neither [1], [2]. See [3], [4] for discussions of wavelet expansions using nonindependent wavelets, with continuous [3] or discrete [4] dilation and translation labels. Even the special case of orthonormal wavelets need not always be of the form (1.1). Basic dilation factors different from 2 are possible: there exist orthonormal bases in which this factor is any rational $p/q > 1$ [5]; in more than one dimension we may even choose a dilation *matrix* instead of an isotropic dilation factor. In these more general cases, it may be necessary to introduce more than one ψ (but always a finite number). We shall restrict ourselves to one dimension here, and to the dilation factor 2, as in (1.1). Bases with factor 2 are by far the easiest to implement for numerical computations.

All interesting examples of orthonormal wavelet bases can be constructed via *multiresolution analysis.* This is a framework developed by Mallat [6] and Meyer [7], in which the wavelet coefficients $\langle f, \psi_{jk}\rangle$ for fixed j describe the difference between two approximations of f, one with resolution 2^{j-1}, and one with the coarser resolution 2^j. The following succinct review of multiresolution analysis suffices for the understanding of this paper; for more details, examples, and proofs we refer the reader to [6] and [7].

The successive approximation spaces V_j in a multiresolution analysis can be characterized by means of a *scaling function* ϕ. More precisely, we assume that the integer translates of ϕ are an orthonormal basis for the space V_0, which we define to be the approximation space with resolution 1. The approximation spaces V_j with resolution 2^j are then defined as the closed linear spans of the ϕ_{jk} $(k \in \mathbb{Z})$, where

$$\phi_{jk} = 2^{-j/2}\phi(2^{-j}x - k). \tag{1.2}$$

To ensure that projections on the V_j describe successive approximations, we require $V_0 \subset V_{-1}$, which implies

$$\cdots V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \cdots. \tag{1.3}$$

* Received by the editors May 29, 1990; accepted for publication (in revised form) May 23, 1992.

† Mathematics Department, Rutgers University, New Brunswick, New Jersey 08903 and AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974.

This imposes a restriction on ϕ: since $\phi \in V_0 \subset V_{-1} = \overline{\operatorname{Span}\{\phi_{-1k};\, k \in \mathbb{Z}\}}$, there must exist c_n such that

$$\phi(x) = \sum_n c_n \phi(2x - n). \tag{1.4}$$

In order to have a complete description of $L^2(\mathbb{R})$, we also impose

$$\bigcap_{j \in \mathbb{Z}} V_j = \{0\}, \qquad \overline{\bigcup_{j \in \mathbb{Z}} V_j} = L^2(\mathbb{R}). \tag{1.5}$$

For every multiresolution analysis as described above, there exists a corresponding orthonormal basis of wavelets defined by

$$\psi(x) = \sum_{n \in \mathbb{Z}} (-1)^n c_{-n+1} \phi(2x - n), \tag{1.6}$$

where c_n are the coefficients in (1.4). We can prove [6], [7] (see also below) that the ψ_{0n} are then an orthonormal basis for the orthogonal complement W_0 of V_0 in V_{-1}. This phenomenon repeats itself at every resolution level j. It follows that, for every j, the $\langle f, \psi_{jk} \rangle$ determine the difference in information between the approximations $P_j f$, $P_{j-1} f$ at resolutions 2^j, 2^{j-1}, respectively:

$$P_{j-1} f = P_j f + \sum_k \langle f, \psi_{jk} \rangle \psi_{jk}.$$

Consequently, by (1.3) and (1.5), the $(\psi_{jk};\, j, k \in \mathbb{Z})$ constitute an orthonormal basis for $L^2(\mathbb{R})$.

One advantage of the "nested" structure of a multiresolution analysis is that it leads to an efficient tree-structured algorithm for the decomposition and reconstruction of functions (given either in continuous or sampled form). Instead of computing all the inner products $\langle f, \psi_{jk} \rangle$ directly, we proceed in a hierarchic way:

—compute $\langle f, \phi_{jk} \rangle$ for the finest resolution level j wanted (if the data are given in a discrete fashion, then these discrete data can just be taken to be $\langle f, \phi_{jk} \rangle$);

—then compute $\langle f, \psi_{j-1k} \rangle$ *and* $\langle f, \phi_{j-1k} \rangle$ at the next finest resolution level by applying (1.4) and (1.7),

$$\langle f, \psi_{j-1k} \rangle = \frac{1}{\sqrt{2}} \sum_n (-1)^n c_{-n+2k+1} \langle f, \phi_{jn} \rangle,$$

$$\langle f, \phi_{j-1k} \rangle = \frac{1}{\sqrt{2}} \sum_n c_{n-2k} \langle f, \phi_{jn} \rangle;$$

—iterate until the coarsest desired resolution level is attained.

The total complexity of this calculation is lower, despite the computation of the seemingly unnecessary $\langle f, \phi_{jk} \rangle$, than if the $\langle f, \psi_{jk} \rangle$ were computed directly.

This brief review shows how to construct an orthonormal basis of wavelets from any "decent" function ϕ satisfying an equation of type (1.4). An example of such a construction is given by the Battle–Lemarié wavelets, consisting of spline functions [8], [9], [10]. In general, constructions starting from a choice of ϕ lead to ϕ, ψ, which are not compactly supported (see, e.g., [15], [25] for a more detailed discussion). The construction can, however, also be viewed differently. The Fourier transform of (1.4) is

$$\hat{\phi}(\xi) = \left[\frac{1}{2} \sum_n c_n e^{in\xi/2}\right] \hat{\phi}\left(\frac{\xi}{2}\right),$$

which implies

$$\hat{\phi}(\xi) = \left[\prod_{j=1}^{\infty} m_0(2^{-j}\xi)\right] \hat{\phi}(0), \tag{1.7}$$

with $m_0(\xi)=\frac{1}{2}\sum_n c_n e^{in\xi}$, so that, up to normalization, ϕ is completely determined by the c_n. Fixing the c_n, therefore, also defines a multiresolution analysis. The c_n have to satisfy certain conditions. Combining $\langle\phi_{0k},\phi_{0l}\rangle=\delta_{kl}$ with (1.4) immediately leads to

$$(1.8)\qquad \sum_n c_n c_{n-2k}=2\delta_{k0},$$

where we have assumed, as we shall do in the sequel, that the c_n are real. In terms of $m_0(\xi)$, (1.8) can be rewritten as

$$(1.9)\qquad |m_0(\xi)|^2+|m_0(\xi+\pi)|^2=1.$$

To ensure that ϕ is well defined, the infinite product in (1.7) must converge, which implies $m_0(0)=1$ or

$$(1.10)\qquad \sum_n c_n=2.$$

It follows that ϕ is uniquely determined by (1.4), up to normalization, which we fix by requiring $\int dx\,\phi(x)=1$. One can show (see, e.g., [12]) that (1.9) implies that ϕ is in $L^2(\mathbb{R})$, but unfortunately (1.8) is not sufficient to guarantee orthonormality of the ϕ_{0n}. A counterexample is $c_0=c_3=1$, all other $c_n=0$, which leads to $\phi(x)=\frac{1}{3}$ for $0\leqq x<3$, $\phi(x)=0$ otherwise. Such counterexamples are rare, however. If $N_2-N_1=3$, then the example above, $c_0=c_3=1$, is the only one. For a detailed discussion, see [12], [13], [22].

If we exclude these thin sets of "bad" choices for the c_n (which can be done by various means [6], [7], [12], [13], [15]), then we can build orthonormal bases of wavelets starting from the c_n. Once orthonormality of the ϕ_{0k} is established, all the rest follows easily. Formula (1.6) for ψ leads immediately to orthogonality of the ψ_{0l} and ϕ_{0k},

$$\begin{aligned}\langle\psi_{0l},\phi_{0k}\rangle&=\frac{1}{2}\sum_{n,m}(-1)^n c_{-n+2l+1}c_{m-2k}\langle\phi_{-1n},\phi_{-1m}\rangle\\&=\frac{1}{2}\sum_n(-1)^n c_{-n+2l+1}c_{n-2k}=0.\end{aligned}$$

The last equality follows from the substitution $n=-m+2(k+l)+1$ for the summation index n. Similar manipulations prove

$$\langle\psi_{0l},\psi_{0k}\rangle=\delta_{kl}$$

and

$$(1.11)\qquad \sum_k[\langle f,\phi_{0k}\rangle\phi_{0k}+\langle f,\psi_{0k}\rangle\psi_{0k}]=\sum_n\langle f,\phi_{-1n}\rangle\phi_{-1n}.$$

It follows that both $\{\phi_{-1n};\,n\in\mathbb{Z}\}$ and $\{\phi_{0k},\psi_{0k};\,k\in\mathbb{Z}\}$ are orthonormal bases for V_{-1}. (In other words, (1.8) ensures that (1.4) and (1.6) describe an orthonormal basis transformation.) It follows that $W_0=\overline{\text{Span}\,(\psi_{0k})}$ is the orthogonal complement of V_0 in V_{-1}, and hence that the $\{\psi_{jk};\,j,k\in\mathbb{Z}\}$ constitute an orthonormal basis for $L^2(\mathbb{R})$.

Constructing ψ from the c_n rather than from ϕ has the advantage of allowing better control over the supports of ϕ and ψ. If $c_n=0$ for $n<N_1$, $n>N_2$, then support $(\phi)\subset[N_1,N_2]$ (see [11a], [14]). In [15] this method was used to construct orthonormal bases of wavelets with compact support, and arbitrarily high preassigned regularity (the size of the support increases linearly with the number of continuous derivatives). These orthonormal basis functions and the associated multiresolution analysis have been tried out for several applications, ranging from image processing to numerical analysis [16]. For some of these applications, variations on the scheme of [15] were requested, emphasizing other properties. The goal of this and the next paper is to present a number of these variations.

The construction in [15] relied on the identity

$$\sum_{j=0}^{N-1}\binom{N-1+j}{j}[(\cos\alpha)^{2N}(\sin\alpha)^{2j}+(\sin\alpha)^{2N}(\cos\alpha)^{2j}]=1. \tag{1.12}$$

Since

$$\left|\frac{1+e^{i\xi}}{2}\right|^2=\left(\cos\frac{\xi}{2}\right)^2,$$

(1.12) suggests the choice

$$m_0(\xi)=\left(\frac{1+e^{i\xi}}{2}\right)^N Q(e^{i\xi}), \tag{1.13}$$

where Q is a trigonometric polynomial with real coefficients such that

$$|Q(e^{i\xi})|^2=\sum_{j=0}^{N-1}\binom{N-1+j}{j}\left(\frac{1-\cos\xi}{2}\right)^j. \tag{1.14}$$

By (1.12), any such m_0 will satisfy (1.9). To determine Q, we have to extract the "square root" of the right-hand side of (1.5). This can be done by using a lemma of Riesz [17]. Denote the right-hand side of (1.14) by $P_N(e^{i\xi})$, and extend P_N to all of $\mathbb{C}$. We have $\overline{P_N(z)}=P_N(\bar{z})$ and $P_N(z^{-1})=P_N(z)$. Consequently, the zeros of P_N come either in real duplets, r_k and r_k^{-1}, or in complex quadruplets, z_l, $\bar{z}_l$, z_l^{-1}, and $\bar{z}_l^{-1}$,

$$\begin{aligned}P_N(z)&=4^{-N+1}\binom{2N-2}{N-1}z^{-N+1}\prod_k(z-r_k)(z-r_k^{-1})\\&\quad\cdot\prod_l(z-z_l)(z-\bar{z}_l)(z-z_l^{-1})(z-\bar{z}_l^{-1})\\&=4^{-N+1}\binom{2N-2}{N-1}\prod_k\frac{(z-r_k)(r_k-z^{-1})}{r_k^2}\\&\quad\cdot\prod_l\frac{(z-z_l)(z-\bar{z}_l)(z_l-z^{-1})(\bar{z}_l-z^{-1})}{|z_l|^2z^2}.\end{aligned}$$

It follows that $P_N(e^{i\xi})=|Q_N(e^{i\xi})|^2$, with

$$Q_N(z)=2^{-N+1}\binom{2N-2}{N-1}^{1/2}\prod_k\frac{(z-r_k)}{\sqrt{|r_k|}}\prod_l\frac{(z^2+|z_l|^2-2z\operatorname{Re}z_l)}{|z_l|}. \tag{1.15}$$

This gives a recipe for the construction of m_0:

(1) For given N, determine the zeros of P_N;

(2) Choose one zero out of every pair of real zeros r_l, r_l^{-1} of P_N, and one conjugated pair out of every quadruplet z_k, z_k^{-1}, $\bar{z}_k$, $\bar{z}_k^{-1}$.

(3) Compute the product Q_N, and substitute into (1.12).

The result is a polynomial in $e^{i\xi}$ of degree $2N-1$, corresponding to an orthonormal basis of wavelets in which the basic wavelet ψ_N has support width $2N-1$. Since (1.6) can be rewritten as

$$\hat{\psi}(\xi)=e^{i((\xi/2)+\pi)}\overline{m_0\left(\frac{\xi}{2}+\pi\right)}\hat{\phi}\left(\frac{\xi}{2}\right),$$

and since (1.13) has a zero of order N at π, it follows that ψ_N has N vanishing moments,

$$\int dx\, x^l \psi_N(x) = 0, \qquad l = 0, 1, \ldots, N-1,$$

which is useful for quantum field theory [18] and numerical analysis applications [19]. The regularity of the ψ_N constructed in [15] increases linearly with their support width, $\psi_N \in C^{\alpha(N)}$, with $\lim_{N\to\infty} N^{-1}\alpha(N) = .2075$ [23], [24], [25]. Plots of ϕ and ψ for various values of N can be found in [15], [25].

Depending on the application they had in mind, several scientists (mathematicians or engineers) have requested possible variations on the construction in [15]. The following are the most recurrent wish items.

(1) More symmetry: the functions ϕ, ψ in [15] are very asymmetric. Complete symmetry is incompatible with the orthonormal basis condition (see [15, p. 971], or § 2 below), but is less asymmetry possible?

(2) Better frequency resolution: orthonormal bases with basic multiplication factor 2 correspond to frequency intervals of 1 octave. Is better possible (e.g., $\frac{1}{2}$ octave), without giving up compact support?

(3) More regularity: is better regularity than in [15] achievable for the same support width?

(4) More vanishing moments: for a fixed support width $2N-1$, the ψ_N of [15] have the maximum number of vanishing moments. The functions ϕ_N do not satisfy any moment condition, except $\int dx\, \phi_N(x) = 1$. For numerical analysis applications, it may be useful to give up some zero moments of ψ in order to obtain zero moments for ϕ, i.e., to have

$$\begin{aligned} &\int dx\, \phi(x) = 1, \\ (1.16) \qquad &\int dx\, x^l \phi(x) = 0, \qquad l = 1, \ldots, L, \\ &\int dx\, x^l \psi(x) = 0, \qquad l = 0, \ldots, L. \end{aligned}$$

How can such ϕ, ψ be constructed? They would have the advantage that inner products with smooth functions are particularly appealing:

$$\begin{aligned} \int dx\, \phi_{-jk}(x) f(x) &= 2^{j/2} \int dx\, \phi(2^j(x - 2^{-j}k)) f(x) \\ &= 2^{-j/2} f(2^{-j}k) + \text{correction terms in } f^{(L+1)} \end{aligned}$$

(use the Taylor expansion of f around $2^{-j}k$; the second through $(L+1)$th terms vanish because of (1.16)). Moreover, if the $(L+1)$th derivative of f is uniformly bounded, then the correction terms in this formula are of order $2^{-(L+1/2)j}$.

The purpose of this and the next paper is to show how such variations can be constructed. In § 2 we handle symmetry, in § 3 regularity, and in § 4 vanishing moments for ϕ. The next paper shows how to obtain better frequency localization.

2. More symmetry. If we restrict our attention to orthonormal bases of *compactly supported wavelets only*, then it is impossible to obtain ψ which is either symmetric or antisymmetric, except for the trivial Haar case ($c_0 = 1$, $c_1 = -1$, all other $c_n = 0$). This is the content of the following theorem.

THEOREM 2.1. *Let ψ, ϕ be defined as in* § 1, *from a finite set of coefficients* c_n *satisfying* (1.9) *and* (1.11), *with orthonormal* ϕ_{0n}. *If* ψ *is either symmetric or antisymmetric around some axis, then* ψ *is the Haar function.*

A proof can be found in [25, Chap. 8].

It is thus a fact of life that symmetric or antisymmetric ψ, however desirable they might be in applications, are just not possible within a framework of orthonormal bases of continuous, compactly supported wavelets. On the other hand, ϕ and ψ do not really need to be quite as asymmetric as in [15], where the extreme asymmetry of ψ, ϕ proceeds from choices made in their construction. In practice, the $2(N-1)$ zeros of P_N consist of one real pair r, r^{-1} and n_0^{-1} quadruplets of complex zeros z_l, $\bar{z}_l$, z_l^{-1}, $\bar{z}_l^{-1}$ if $N=2n_0$ is even, and of n_0 quadruplets if $N=2n_0+1$ is odd. To construct Q_N, we need to select one of the two real zeros, and one pair z_l, $\bar{z}_l$ out of every quadruplet. The choice made in [15] is the so-called *extremal phase* choice: we chose systematically all zeros with modulus smaller than one. Other choices may lead to less asymmetric ϕ. The following argument shows why.

A sequence of real numbers $(\alpha_n)_{n\in\mathbb{Z}}$ is said to define a *linear phase filter* if the phase of the function $\alpha(\xi)=\sum_n \alpha_n e^{in\xi}$ is a linear function of ξ, i.e., if, for some $l\in\mathbb{Z}/2$,

$$\alpha(\xi)=e^{il\xi}|\alpha(\xi)|.$$

This means that the α_n are symmetric around l, $\alpha_n=\alpha_{2l-n}$. If the sequence does not define a linear phase filter, then the deviation from linearity of the phase of $\alpha(\xi)$ reflects the asymmetry of the α_n. The Fourier transform of ϕ is given by the infinite product (1.7). If c_n were symmetric around l, then we would have $m_0(\xi)=e^{il\xi}|m_0(\xi)|$, hence

$$\hat{\phi}(\xi)=\exp\left[il\sum_{j=1}^{\infty}2^{-j}\xi\right]\prod_{j=1}^{\infty}|m_0(2^{-j}\xi)||\hat{\phi}(0)|=e^{il\xi}|\hat{\phi}(\xi)|,$$

so that ϕ would be symmetric around l as well. As explained above, this is impossible for c_n satisfying (1.8). The closer the phase of m_0 is to linear phase, the closer the phase of $\hat{\phi}$ will be to linear phase, and the less asymmetric ϕ will be. In our case, m_0 is a product of factors of type

$$(2.1)\qquad \begin{aligned}(z-z_l)(z-\bar{z}_l)&=e^{i\xi}(e^{i\xi}-R_l\,e^{i\alpha_l})(1-e^{-i\xi}R_l\,e^{-i\alpha_l})\\ &=e^{i\xi}[e^{i\xi}-2R_l\cos\alpha_l+R_l^2\,e^{-i\xi}],\end{aligned}$$

with possibly an extra factor

$$(2.2)\qquad (z-r)=e^{i\xi/2}[e^{i\xi/2}-r\,e^{i\xi/2}].$$

The total phase of m_0 is a sum of the phase contributions of each factor. Apart from linear phase terms, the phase contributions of (2.1) and (2.2) are, respectively,

$$(2.3)\qquad \Phi_l(\xi)=\text{arctg}\left(\frac{(1-R_l^2)\sin\xi}{(1+R_l^2)\cos\xi-2R_l\cos\alpha_l}\right),$$

$$(2.4)\qquad =\text{arctg}\left(\frac{1+r}{1-r}\,tg\frac{\xi}{2}\right).$$

The valuation of arctg should be chosen so that Φ_l is continuous in $[0,2\pi]$, and $\Phi_l(0)=0$. Since the denominator in (2.3) has two zeros, namely,

$$\xi=\xi_l=\text{Arc}\cos\left(\frac{2R_l}{1+R_l^2}\cos\alpha_l\right)$$

and $\xi = 2\pi - \xi_l$, $\Phi_l(2\pi) = \Phi_l(0) + \varepsilon_l 2\pi$, with $\varepsilon_l = \pm 1$. Something similar happens in the $(z-r)$ case. In order to extract only the nonlinear part of Φ_l, we define, therefore,

$$\Psi_l(\xi) = \operatorname{arctg}\left(\frac{(1-R_l^2)\sin\xi}{(1+R_l^2)\cos\xi - 2R_l\cos\alpha_l}\right) - \frac{\xi}{2\pi}\Phi_l(2\pi)$$

or

$$= \operatorname{arctg}\left(\frac{1+r}{1-r}\, tg\frac{\xi}{2}\right) - \frac{\xi}{2\pi}\Phi_l(2\pi).$$

In order to obtain m_0 as close to linear phase as possible, we have to choose the zeros to retain from every quadruplet or duplet in such a way that $\Psi_{tot}(\xi) = \sum_l \Psi_l(\xi)$ is as close to zero as possible. In practice, we have $2^{\lfloor N/2 \rfloor}$ choices (and not 2^{N-1}, as was mistakenly stated in [15]). This number can be reduced by another factor of 2: for every choice, the complementary choice (choosing all the other zeros) leads to the complex conjugate m_0 (up to a phase shift), and, therefore, to the mirror image of ϕ. For $N = 2$ or 3, there is, therefore, effectively only one pair ϕ_N, ψ_N. For $N \geqq 4$, we can compare the $2^{\lfloor N/2 \rfloor - 1}$ different choices for Ψ_{tot} in order to find the closest to linear phase. It turns our that the net effect of a change of choice from z_l, $\bar{z}_l$ to z_l^{-1}, $\bar{z}_l^{-1}$ is most significant if R_l is close to 1, and if α_l is close to either zero or π. In Fig. 1 we show the graphs for $\Psi_{tot}(\xi)$ for $N = 4$ and 10, both for the original construction in [15] and for the case with flattest Ψ_{tot}. The "least asymmetric" ϕ and ψ, associated with the flattest possible Ψ_{tot}, are plotted in Fig. 2 for $N = 4$ and 10. A table for the corresponding c_n can be found in [25, p. 198], as well as figures for $N = 6, 8$.

Remarks.

(1) In this discussion we have restricted ourselves to the case where m_0 and $|Q|^2$ are given by (1.13) and (1.14), respectively. This means that the ϕ in Fig. 2 are the least asymmetric possible, given that N moments of ψ are zero, and that ϕ has support width $2N-1$. (This is the minimum width for N vanishing moments.) If ϕ may have larger support width, then it can be made even more symmetric. These wider solutions correspond to a variation on (1.14), i.e., to

$$|Q(e^{i\xi})|^2 = \sum_{j=0}^{N-1} \binom{N-1+j}{j} \left(\frac{1-\cos\xi}{2}\right)^j + \left(\frac{1-\cos\xi}{2}\right)^N R(\cos\xi), \tag{2.5}$$

where R is any odd polynomial such that the right-hand side of (2.5) is positive for

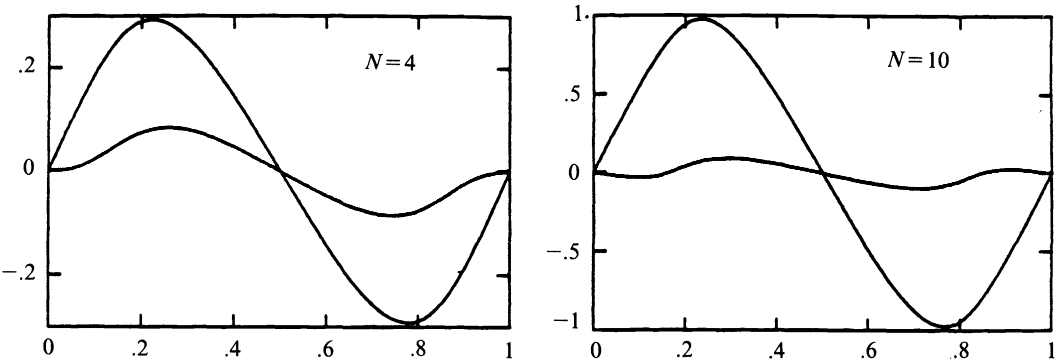


FIG. 1. *Plots of $\Psi_{tot}(\xi)$ for the cases $N = 4$ and 10. In both cases we plot Ψ_{tot} for the construction in [15], and the much flatter Ψ_{tot} corresponding to the closest to linear phase choice. The horizontal axis gives $\xi/2\pi$, the vertical axis Ψ_{tot}/π.*

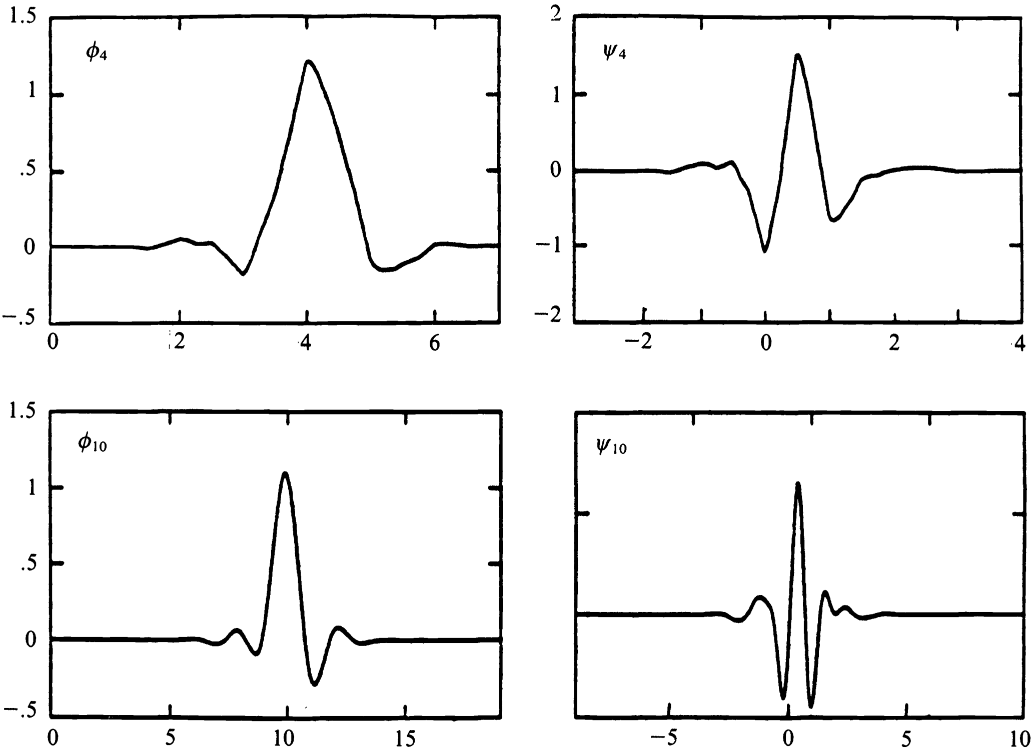


FIG. 2. *Plots of* ϕ_N, ψ_N *closest to linear phase, for the cases* $N=4$ *and* 10. *In every case, support* $(\phi_N)=[0, 2N-1]$, *support* $(\psi_N)=[-N+1, N]$.

all ξ. The functions ϕ constructed in § 4, for instance, are more symmetric than those in Fig. 2, but they have large support width.

(2) We can achieve even more symmetry by going a little beyond the multiresolution scheme explained in § 1, and by "mirroring" the filters at every odd step. For more details, see [25, p. 256].

(3) In [21] the construction of orthonormal bases of wavelets is generalized to "biorthogonal bases," i.e., to two dual unconditional bases $\{\psi_{jk};\ j, k\in\mathbb{Z}\}$ and $\{\tilde{\psi}_{jk};\ j, k\in\mathbb{Z}\}$. The construction in [21] corresponds to a decomposition+reconstruction scheme in which the reconstruction filters differ from the decomposition filters. In this more general framework, complete symmetry *can* be achieved. Orthonormality is then lost, however, which is less desirable for some applications.

3. More regularity. The regularity of the wavelets ψ_n constructed in [15] increases linearly with their support width, $\psi_N\in C^{\alpha(N)}$, $\lim N^{-1}\alpha(N)=.2075$. The technique used in [15] to control the regularity of ϕ_N, ψ_N involved constructing $m_0(\xi)$ so that it contained the factor $\frac{1}{2}(1+e^{i\xi})$ with as high multiplicity as possible,

$$m_0(\xi)=\left(\frac{1+e^{i\xi}}{2}\right)^N Q_N(\xi), \tag{3.1}$$

where Q_N is a polynomial in $e^{i\xi}$ of order $N-1$ (see § 1). Since $\prod_{j=0}^{\infty}(1+\exp(i2^{-j}\xi))/2=e^{i\xi}(\sin\xi/\xi)$, we find (use (1.7))

$$\hat{\phi}_N(\xi)=e^{iN\xi/2}\left[\frac{\sin\xi/2}{\xi/2}\right]^N\prod_{j=1}^{\infty}Q_N(2^{-j}\xi).$$

Together with control on the infinite product of Q_N (see [15]), this leads to decay for $\hat{\phi}_N$ as $|\xi|\to\infty$, hence to regularity for ϕ_N, ψ_N.

In this argument, imposing high order divisibility of m_0 by $\frac{1}{2}(1+e^{i\xi})$ is used as a technical tool to obtain regularity. On the other hand, regularity for ϕ *implies* that m_0 is of type (3.1). More precisely, if ϕ is compactly supported and $\phi\in C^L$, then m_0 must be divisible by $[\frac{1}{2}(1+e^{i\xi})]^L$; see [22], [21]. Since $\phi_N\in C^{\mu N}$ for large N, with $\mu\simeq .2$, this means that at least $\frac{1}{5}$ of the factors $(1+e^{i\xi})$ in $m_{0,N}$ are necessary. Can the others be dispensed with, allowing even shorter support for the same regularity, or higher regularity for the same support width? The answer is yes.

In [11b], an alternative way was used to determine the regularity of functions ϕ satisfying an equation of type (1.4). Unlike the methods in [15], the method of [11b] does not use the Fourier transform. Instead, two N-dimensional matrices T_0, T_1 are defined, $(T_0)_{i,j}=c_{2i-j-1}$, $(T_1)_{i,j}=c_{2i-j}$, $1\leqq i, j\leqq N$, where we assume $c_n=0$ for $n<0$ or $n>N$. Divisibility of m_0 by $(1+e^{i\xi})$ with multiplicity L is equivalent to

$$\sum_{n=0}^{N} c_n(-1)^n n^l=0, \qquad l=0,\ldots,L-1. \tag{3.2}$$

In terms of the matrices T_0, T_1, this implies that there exists a flag of subspaces $U_1\subset\cdots\subset U_L$ of $\mathbb{R}^N$, with $\dim U_j=j$, such that

- U_j is left-invariant under both T_0, T_1.
- The left restrictions of T_0, T_1 to U_j have the j eigenvalues $1,\frac{1}{2},\ldots,2^{-j+1}$.

Let V_L be the subspace for $\mathbb{R}^N$ orthogonal to U_L; V_L is right invariant for T_0, T_1. If, for some $\lambda<1$, $C>0$, and for all $m\in\mathbb{N}$,

$$\|T_{d_1}\cdots T_{d_m}|_{V_L}\|\leqq C\lambda^m 2^{-m(L-1)} \qquad (d_j=1 \text{ or } 0), \tag{3.3}$$

then (3.2) implies that $\phi\in C^L$, and that its Lth derivative $\phi^{(L)}$ is Hölder continuous with exponent $|\log_2\lambda|$; if λ is best possible in (3.3), then $|\log_2\lambda|$ is the best possible Hölder exponent for $\phi^{(L)}$. In principle (3.3) involves infinitely many inequalities; in practice we substitute finitely many conditions sufficient to ensure that (3.3) holds for all m [11b, Prop. 3.11]. The value of ϕ and its derivatives at any point x in support (ϕ) is governed by the behavior of the infinite product $T_{d_1(x)}T_{d_2(x)}\cdots T_{d_m(x)}\cdots$, where $d_j(x)$ are the digits in the binary expansion of x, $x=\lfloor x\rfloor+\sum_{j=1}^{\infty} d_j(x)2^{-j}$. Special, "local" inequalities of type (3.3), valid only for certain sequences $(d_n)_{n\in\mathbb{N}}$, can, therefore, be translated into local regularity estimates, leading, in many examples, to a hierarchy of fractal sets corresponding to different local Hölder exponents. For more details, see [11b].

This approach can be used to study the regularity of compactly supported basis wavelets, which all correspond to an equation of type (1.4) with finitely many coefficients. For the examples of [15], this analysis was carried out in [11b] for $N=2,3,4$ (for higher N, checking (3.3) becomes very complicated). In these three cases, the best possible Hölder exponent for the highest order well-defined derivative of ϕ_N was determined; these results were significantly better than what had been obtained in [15] via Fourier analysis. Table 1 compares the regularity results of [15] and [11b].

The optimal estimates obtained in [11] illustrate again that some of the factors $(1+e^{i\xi})$ of m_0, or, equivalently, some of the sum rules (3.2), which we impose in order to obtain regularity, are "wasted" in the final construction. N sum rules can deliver up to $N-1$ continuous derivatives if everything else cooperates; because of the other constraints on the c_n (i.e., (1.8)), wavelets do not achieve this optimal number. We can, therefore, drop some of the sum rules, and use the additional degrees of freedom

TABLE 1
The regularity of the wavelets of [15], *as obtained via Fourier methods (middle column) or via the matrix method of* [11b] *(right column). The integer part of the entry is the number of times* ϕ *is continuously differentiable; the decimal part is the Hölder exponent of the highest order well-defined derivative.*

N	Best estimate in [15]	Optimal result, obtained in [11b]
2	$.5-\varepsilon$	.5500...
3	.915	1.0878...
4	1.275	1.5179...

to obtain better λ in (3.3), i.e., better regularity than in Table 1. We present here the cases of wavelets with support width 3 and 5.

For $|\text{support } \phi|=3$, there is a one-parameter family of choices c_n satisfying (1.8) and (1.10) (see [15, p. 946]), namely,

$$c_0=\frac{\nu(\nu-1)}{(\nu^2+1)},\quad c_1=\frac{(1-\nu)}{(\nu^2+1)},\quad c_2=\frac{(\nu+1)}{(\nu^2+1)},\quad c_3=\frac{\nu(\nu+1)}{(\nu^2+1)}.$$

These c_n satisfy $\sum c_n(-1)^n=0$; imposing a second sum rule leads to $\nu=\pm 3^{-1/2}$, which corresponds to the "standard" case $N=2$. The matrices T_0 and T_1 are 3×3-matrices; since we have one sum rule, we can restrict our attention to the reduced matrices $T_0|_{V_1}$, $T_1|_{V_1}$,

$$T_0|_{V_1}=\frac{1}{\nu^2+1}\begin{pmatrix}\nu(\nu-1) & 0\\ 1-\nu^2 & \nu(\nu+1)\end{pmatrix},\qquad T_1|_{V_1}=\frac{1}{\nu^2+1}\begin{pmatrix}\nu(\nu+1) & \nu(\nu-1)\\ 0 & 1-\nu^2\end{pmatrix}.$$

We restrict our attention to $\nu\geqq 0$. (A change of sign $\nu\to-\nu$ corresponds to $c_n\to c_{3-n}$, i.e., to mirroring ϕ with respect to $\frac{3}{2}$.) Since (3.3) has to hold, in particular when all the d_j are identical, $d_j\equiv 0$ or $d_j\equiv 1$, the constant λ is bounded below by the spectral radii $\rho(T_j|_{V_1})$ of $T_j|_{V_1}$, $j=0$ or 1. It follows that (3.3) can only be satisfied if $\nu<1$. For $\nu\geqq 1/\sqrt{3}$, we can find M so that both $MT_j|_{V_1}M^{-1}$, $j=0,1$, are symmetric; consequently,

$$\begin{aligned}\lambda &\leqq \max(\|MT_j|_{V_1}M^{-1}\|;\ j=0,1)=\max(\rho(MT_j|_{V_1}M^{-1});\ j=0,1)\\ &=\max(\rho(T_j|_{V_1});\ j=0,1)=\frac{\nu(\nu+1)}{1+\nu^2}.\end{aligned}$$

This is, moreover, the best possible λ. If $\nu<1/\sqrt{3}$, then $T_0|_{V_1}$ and $T_1|_{V_1}$ are not simultaneously "symmetrizable," and we have to do some more work. In every case

$$\lambda\geqq\max\left(\frac{\nu(\nu+1)}{1+\nu^2},\ \frac{1-\nu^2}{1+\nu^2}\right). \tag{3.4}$$

For $\nu=.25$, e.g., the tricks of Proposition 3.11 in [11b] suffice to show that equality holds in (3.4), and

$$\lambda=\frac{1-1/16}{1+1/16}=.88235\ldots.$$

The lowest value for the right-hand side of (3.4), and, therefore, the best candidate for the "most regular possible" ϕ, occurs for $\nu=.5$. In this case, $T_1|_{V_1}$ has only one (degenerate) eigenvalue, .6, and the matrix is not diagonalizable. Since (3.3) has to hold for $d_j\equiv 1$, it follows that we can at best hope to establish $\lambda=.6(1+\varepsilon)$.

In fact, we cannot achieve even this much. It turns out that $[\rho(T_0, T^{12})]^{1/13} \simeq .659676 \cdots > .6$, meaning that we can certainly not hope for a smaller λ than $.659 \cdots$. Using all the tricks in Proposition 3.11 in [11b], and checking a collection of building blocks with up to 17 factors, we find $\lambda \leqq .666$. More work leads to smaller upper bounds for λ; presumably the best value is the .659 obtained above.

Figure 3 shows the function ϕ for a few choices of ν ($\nu = .75$, .5 and .25). In each case ϕ is continuous, and we can compute its Hölder exponent from our estimate for λ. Even with our less than optimal estimate $\lambda \leqq .666$, the case $\nu = .5$ leads to a better Hölder exponent than the "standard" example $\nu = 1/\sqrt{3}$. This might be surprising: the graph of ϕ for $\nu = .5$ seems more jagged than for $\nu = 1/\sqrt{3}$. However, the peaks in the $\nu = .5$ example are "less sharp": the steepest slope of the peak around $x = 1$, e.g., is

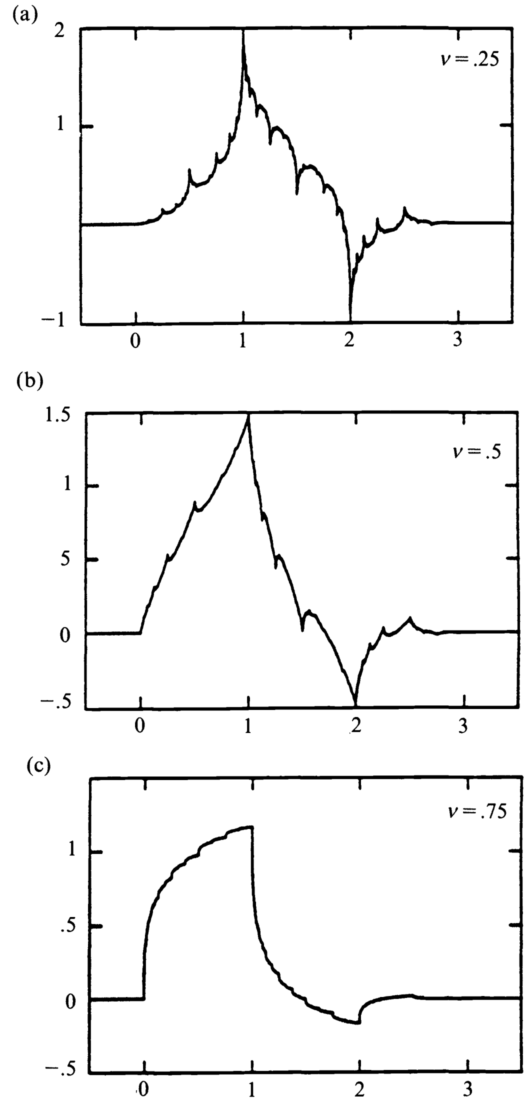


FIG. 3. *The functions ϕ defined by $c_0 = \nu(\nu+1)/(1+\nu^2)$, $c_1 = (\nu+1)/(\nu^2+1)$, $c_2 = (1-\nu)/(\nu^2+1)$, $c_3 = \nu(\nu-1)/(\nu^2+1)$, for different values of ν. As outlined in the text we can prove that the Hölder exponents of these functions are at least* (a) .180572, (b) .5864, (c) .251539. *For* a, c *these numbers are sharp: for* b *the true Hölder exponent is conjectured to be* .60017.

less steep than its counterpart for $\nu = 1/\sqrt{3}$, and this steepness is what is really expressed by a low Hölder exponent.

For |support ϕ| $= 5$ we have no analytical expression for all the possible choices of the c_n. Since the "standard" example, with its 3 sum rules, achieves C^1-regularity (see Table 1), for which at least 2 sum rules are necessary, we can drop at most one sum rule. We explore what this extra degree of freedom can give us by perturbing around the standard example. More precisely, we have

$$m_0(\xi) = \left(\frac{1+e^{i\xi}}{2}\right)^2 Q(\xi), \tag{3.5}$$

with

$$|Q(\xi)|^2 = P(\cos\xi), \tag{3.6}$$

$$P(x) = 2 - x + \frac{a}{4}(1-x)^2, \tag{3.7}$$

where a can be chosen freely, subject to the constraint that the right-hand side of (3.6) is nonnegative for all ξ. The example of [15] with support width 5 corresponds to m_0 with a zero of order 3 at $\xi = \pi$, hence to P with a zero at $x = -1$, which gives $a = 3$. If we impose that P has a zero close to $x = -1$, e.g., at $x = -1 - \delta$ (where $\delta \geqq 0$, since otherwise the positivity constraint would be violated), then $a = 4(\delta+3)/(\delta+1)(\delta+2)^2$, and $P(x) = (x+1+\delta)/(\delta+1)(\delta+2)^2[x^2(\delta+3) - x(\delta+3)^2 + 2(\delta+2)^2]$. The other two roots of P are, therefore, given by $x_\pm = \frac{1}{2}(\delta+3) \pm \frac{1}{2}[(\delta+3)^2 - 8(\delta+2)^2/(\delta+3)]^{1/2}$. Each of the three roots of $P(x)$, namely, $x_0 = -1-\delta$, and $x_\pm$, corresponds to two roots in $z = e^{i\xi}$ of $P(\cos\xi)$ (use $\frac{1}{2}(z+z^{-1}) = x \Rightarrow z = x \pm \sqrt{x^2-1}$). This leads to the candidates $Q_\varepsilon(\xi) = N(e^{i\xi} + \delta + 1 + \varepsilon\sqrt{\delta(\delta+2)})\ (e^{i\xi} - z_+(\delta))\ (e^{i\xi} - z_-(\delta))$, where $z_\pm(\delta) = x_\pm(\delta) - \sqrt{x_\pm(\delta)^2 - 1}$ and $\varepsilon = \pm 1$. The choice $\varepsilon = +1$ corresponds to choosing all the zeros of Q inside the unit circle; the choice $\varepsilon = -1$ gives one (real) zero outside, and two complex conjugate zeros inside the unit circle. For $\varepsilon = +1$, the choice $\delta = 0$ (i.e., the example of [15]) minimizes max $(\rho(T_0|_{V_2}), \rho(T_1|_{V_2}))$ (where ρ denotes the spectral radius), so that $\delta = 0$ leads to the most regular ϕ. For $\varepsilon = -1$, the situation is different. We find a minimum for max $(\rho(T_0|_{V_2}), (\rho(T_1|_{V_2}))$ at $\delta = .07645485\ldots$ (value determined numerically). As in the case where |support ϕ| $= 2$, this minimum for the spectral

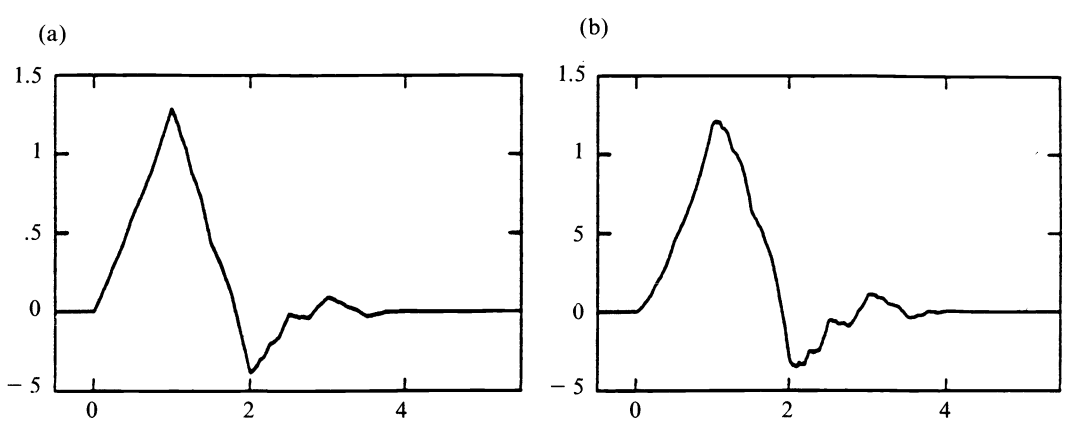


FIG. 4. *Two examples of* ϕ *with* |*support* ϕ| $= 5$. (a) *Corresponds to the construction in* [15], (b) *is the "most regular"* ϕ *constructed here. In both cases* $\phi \in C^1$; *the Hölder exponent of* ϕ' *is* .0878 *for* (a), *and at least* .40198 *for* (b) (*it is conjectured to be* .41762 *for* (b)).

radii $\rho(T_j|_{V_2})$ corresponds to a degenerate largest eigenvalue of $T_1|_{V_2}$, and we find again that $T_1|_{V_2}$ is not diagonalizable. Consequently, we can only hope to establish

$$\lambda \leqq 2(1+\varepsilon)\max\left(\rho(T_0|_{V_2}), \rho(T_1|_{V_2})\right) = (1+\varepsilon).74865\cdots.$$

In order to obtain $\varepsilon < .01$, we already have to consider a large number of building blocks $T_{d_1}\cdots T_{d_m}$, the longest of which has $d_j = 1$ for $j = 1, \ldots, m$, and $m \geqq 700$! It seems likely that arbitrarily small ε can be attained by more work. Figure 4 shows both the standard example of [15] and the most regular ϕ obtained here for $|\text{support } \phi| = 5$. It is apparent that the present example is much more regular; both functions are C^1 (even though the function of [15] seems to have peaks, these peaks are not really sharp—see [11b]), but the Hölder exponent of ϕ' is significantly better in the example constructed here.

4. Vanishing moments for ϕ. In this subsection we want to construct ϕ, ψ with compact support,

$$|\text{supp } \phi| = |\text{supp } \psi| = 2M - 1$$

and such that

$$\int dx\, \phi(x) = 1,$$

$$\int dx\, x^l \phi(x) = 0 \quad \text{for } l = 1, \ldots, L-1, \tag{4.1}$$

$$\int dx\, x^l \psi(x) = 0 \quad \text{for } l = 0, \ldots, L-1.$$

The need for orthonormal bases with this property first came up in the application of wavelet bases to numerical analysis in the work of Beylkin, Coifman, and Rokhlin [19]. The desirability of vanishing moments for ϕ is explained in the introduction: if (4.1) is satisfied, then the inner product of ϕ_{jk} with a smooth function f only depends on $f(2^j k)$ and derivatives of f of order $\geqq L$. (In a later version of their work, Beylkin, Coifman, and Rokhlin did not require (4.1), however.) Imposing such vanishing moments on ϕ also increases its symmetry. Because these orthonormal wavelet bases with vanishing moments for both ϕ and ψ were requested by Coifman, I have named these wavelets *coiflets*. Condition (4.1) corresponds to a *coiflet of order L*.

The Fourier transforms of ϕ, ψ are given by $\hat{\phi}(\xi) = \prod_{j=1}^{\infty} m_0(2^{-j}\xi)$ $\hat{\psi}(\xi) = m_1(\xi/2)\hat{\phi}(\xi/2)$, with

$$m_0(\xi) = \sum_{n=N_1}^{N_2} c_n e^{in\xi}, \qquad m_1(\xi) = \sum_n (-1)^n c_{-n+1} e^{in\xi} = -e^{i\xi}\overline{m_0(\xi+\pi)}.$$

Note that the lower limit N_1 in the sum over n will in general not be zero in this subsection: we have lost our freedom to translate by integers because (4.1) is not invariant under such translations (the conditions on ψ are translation-invariant, but the conditions on ϕ are not). The conditions (4.1) are equivalent to

$$\hat{\phi}(0) = 1, \quad \left(\frac{d^l}{d\xi^l}\hat{\phi}\right)(0) = 0 \quad \text{for } l = 1, \ldots, L-1,$$

$$\left(\frac{d^l}{d\xi^l}\hat{\psi}\right)(0) = 0 \quad \text{for } l = 0, \ldots, L-1.$$

In terms of m_0, these become

$$m_0^{(l)}(\xi+\pi)=0 \quad \text{for } l=0,\dots,L-1, \tag{4.2}$$

$$m_0(0)=1, \quad m_0^{(l)}(0)=0 \quad \text{for } l=1,\dots,L-1. \tag{4.3}$$

By (4.2), m_0 has a zero of order L in $\xi=\pi$. Consequently, m_0 has to be of the form

$$m_0(\xi)=\left(\frac{1+e^{i\xi}}{2}\right)^L Q(e^{i\xi}), \tag{4.4}$$

where

$$|Q(e^{i\xi})|^2=\sum_{j=0}^{L-1}\binom{L-1+j}{j}\left(\frac{1-\cos\xi}{2}\right)^j+\left(\frac{1-\cos\xi}{2}\right)^L R(\cos\xi), \tag{4.5}$$

and R is an odd polynomial [15]. On the other hand (4.3) implies

$$m_0(\xi)=1+(1-e^{i\xi})^L S(e^{i\xi}). \tag{4.6}$$

Together, (4.4) and (4.6) lead to L independent linear constraints on the coefficients of S. Imposing that Q be of the form (4.5), with R an odd polynomial, leads to further quadratic constraints. For small values of L, the whole collection of constraint equations can be solved more or less by hand; for values of L larger than 6, the situation becomes untractable. We propose, therefore, an approach which from the start satisfies (4.2) and (4.3) (the linear constraints on S are built in), and we tackle (4.5) afterwards.

For the sake of convenience, we restrict ourselves to L even, $L=2K$. A similar analysis can be carried out for L odd. We impose that m_0 be of the form

$$m_0(\xi)=\left(\cos^2\frac{\xi}{2}\right)^K\left[\sum_{k=0}^{K-1}\binom{K-1+k}{k}\left(\sin^2\frac{\xi}{2}\right)^k+\left(\sin^2\frac{\xi}{2}\right)^K f(\xi)\right]. \tag{4.7}$$

Since $\cos^2\xi/2=\frac{1}{4}e^{-i\xi}(1+e^{i\xi})^2$, this clearly has a zero of order $2K$ at $\xi=\pi$. On the other hand, (4.7) can be rewritten as (use (1.13))

$$m_0(\xi)=1+\left(\sin^2\frac{\xi}{2}\right)^K\left[-\sum_{k=0}^{K-1}\binom{K-1+k}{k}\left(\cos^2\frac{\xi}{2}\right)^k+\left(\cos^2\frac{\xi}{2}\right)^K f(\xi)\right].$$

This clearly satisfies (4.3). It remains, therefore, to tailor f so that m_0 satisfies (1.10).

For the sake of convenience we shall use f such that

$$f(\xi)=\sum_{n=0}^{K'} f_n e^{in\xi}, \tag{4.8}$$

i.e., $f_n=0$ for all $n<0$. This is by no means the only choice possible; we could also decide to distribute the f_n as symmetrically around zero as possible, so that the support of ϕ would be more symmetrical around $x=0$. It turns out, however, that this symmetrical choice can lead to larger support widths for ϕ than (4.8) (this happens, e.g., for $K=3$). From (4.5) we obtain

$$\begin{aligned}&\left|\sum_{k=1}^{K-1}\binom{K-1+k}{k}\left(\sin^2\frac{\xi}{2}\right)^k+\left(\sin^2\frac{\xi}{2}\right)^K f(\xi)\right|^2\\ &\quad=\sum_{j=0}^{2K-1}\binom{2K-1+j}{j}\left(\sin^2\frac{\xi}{2}\right)^j+\left(\sin^2\frac{\xi}{2}\right)^{2K} R(\cos\xi),\end{aligned} \tag{4.9}$$

where R is an odd polynomial. Rewriting (4.9) leads to

$$\begin{aligned}\left[\sum_{k=0}^{K-1}\binom{K-1+k}{k}s_2^k\right]^2+\sum_{k=0}^{K-1}\binom{K-1+k}{k}s_2^{k+K}[f(\xi)+\overline{f(\xi)}]\\ +s_2^{2K}|f(\xi)|^2=\sum_{j=0}^{2K-1}\binom{2K-1+j}{j}s_2^j+s_2^{2K}R(\cos\xi),\end{aligned}\tag{4.10}$$

where s_2 denotes $\sin^2(\xi/2)$. We shall determine the f_n by identifying coefficients of s_2^j.

Both $f(\xi)+\overline{f(\xi)}$ and $|f(\xi)|^2$ can be written as polynomials in $\cos\xi$, hence in s_2. It follows that only the first term in the left-hand side of (4.10), which is independent of f, contains terms in s_2^j with $j\leqq K-1$. Fortunately, these terms cancel the corresponding terms in s_2^j in the right-hand side of (4.10) because of the identity

$$\sum_{k=0}^{K}\binom{N_1-1+k}{k}\binom{N_2-1+K-k}{K-k}=\binom{N_1+N_2-1+K}{K}.\tag{4.11}$$

(See [26, (5.27)].)

We next concern ourselves with the terms in s_2^j, $j=K,\dots,2K-1$. Only the first two terms in the left-hand side of (4.10) contribute, leading to linear constraints in the f_n. Define g_n by

$$f(\xi)+\overline{f(\xi)}=\sum_{n=0}^{K'}g_ns_2^n.\tag{4.12}$$

Using $s_2=-\frac{1}{4}e^{-i\xi}(1-e^{i\xi})^2$, we find that the f_n and g_n are related through

$$\begin{aligned}f_0&=\frac{1}{2}\sum_{n=0}^{K'}\binom{2n}{n}4^{-n}g_n,\\ f_k&=(-1)^k\sum_{n=k}^{K'}\binom{2n}{n-k}4^{-n}g_n\quad\text{for } k\neq 0.\end{aligned}\tag{4.13}$$

In practice we will determine the g_n and then calculate the f_n and f via (4.13).

Identification of the terms in s_2^j, $j=K,\dots,2K-1$ on both sides of (4.10) gives

$$\begin{aligned}&\sum_{k=j-K+1}^{K-1}\binom{K-1+k}{k}\binom{K-1+j-k}{j-k}\\ &\qquad+\sum_{k=0}^{\min(K',j-K)}\binom{j-1-k}{j-K-k}g_k=\binom{2K-1+j}{j}.\end{aligned}$$

Using (4.11) again, and substituting $j=K+l$, $l=0,\dots,K-1$, we can reduce this to

$$\sum_{m=\max(0,l-K')}^{l}\binom{K-1+m}{m}g_{l-m}=2\sum_{k=0}^{l}\binom{K-1+k}{k}\binom{2K-1+l-k}{K+l-k}.\tag{4.14}$$

This is a system of K linear equations in $\min(K,K'+1)$ unknowns. It has no solutions if $K'+1<K$. If $K'\geqq K-1$, then the invertibility of the triangular matrix

$$M_{ij}=\binom{K-1+i-j}{i-j},\qquad K-1\geqq i\geqq j\geqq 0$$

immediately leads to

$$g_k=2\binom{2K-1+k}{K+k},\qquad k=0,\dots,K-1.$$

It remains to determine the $g_K, \dots, g_{K'}$. They are given by the constraint that

$$\sum_{k=0}^{K-1} \sum_{l=0}^{K'-K} \binom{K-1+k}{k} g_{K+l} s_2^{k+l} + |f(\xi)|^2 \tag{4.15}$$

should be an odd polynomial in $\cos \xi$. Since (4.15) can be rewritten as a polynomial of degree K' in $\cos \xi$, this results in $\lfloor (K'+1)/2 \rfloor$ equations for $K'-K+1$ unknowns. It follows that $K' \geqq 2K-1$ (no miraculous cancellations occur). In the examples worked out here, $K' = 2K-1$. In these examples a solution has to be found for a system of K quadratic equations in K unknowns; every such solution corresponds to a coiflet of order $2K$, with support width $3K-1$.

The system of K equations to be solved can be written out a little more explicitly. Writing x_m, $m = 0, \dots, K-1$ for the K unknown g_{K+m}, we have

$$|f(\xi)|^2 = \sum_{l=-(2K-1)}^{2K-1} e^{il\xi} \sum_{k=\max(0,-l)}^{\min(2K-1, 2K-1-l)} f_k \overline{f_{l+k}}$$

with

$$\begin{aligned} f_k = (1-\tfrac{1}{2}\delta_{k0})(-1)^k \Bigg[2 \sum_{n=k}^{K-1} \binom{2n}{n-k} 4^{-n} \binom{2K-1+n}{K+n} \\ + \sum_{m=0}^{K-1} \binom{2m+2K}{m+K-k} 4^{-m-K} x_m \Bigg], \qquad 0 \leqq k \leqq K-1 \end{aligned} \tag{4.16}$$

$$f_k = (-1)^k \sum_{m=k-K}^{K-1} \binom{2m+2K}{m+K-k} 4^{-m-K} x_m, \qquad K \leqq k \leqq 2K-1. \tag{4.17}$$

On the other hand, the first term in (4.15) can be rewritten as

$$\begin{aligned} &\sum_{j=0}^{2K-2} s_2^j \sum_{m=\max(0,j-K+1)}^{\min(j,K-1)} \binom{K-1+j-m}{j-m} x_m \\ &= \sum_{l=-(2K-2)}^{2K-2} e^{il\xi} (-1)^l \sum_{j=|l|}^{2K-2} 4^{-j} \binom{2j}{j+l} \sum_{m=\max(0,j-K+1)}^{\min(j,K-1)} \binom{K-1+j-m}{j-m} x_m. \end{aligned}$$

The K equations in the unknowns $x_0, \dots, x_{K-1}$ are, therefore,

$$\sum_{k=0}^{2(K-r)-1} f_k \overline{f_{2r+k}} + \sum_{j=2r}^{2K-2} 4^{-j} \binom{2j}{j+2r} \sum_{m=\max(0,j-K+1)}^{\min(j,K-1)} \binom{K-1+j-m}{j-m} x_m = 0, \tag{4.18}$$

where $r = 0, \dots, K-1$, and where (4.16), (4.17) have to be substituted for the f_k.

As a quadratic system (4.18) can have many solutions or no solutions at all. The following heuristic argument suggests that (4.18) will have solutions for sufficiently large K. We can rewrite (4.7) as

$$\begin{aligned} m_0(\xi) = \frac{1}{2} + 2^{-4K+1} K \binom{2K}{K} \sum_{k=0}^{K-1} \frac{(-1)^k}{2k+1} \binom{2K-1}{K+k} (e^{i(2k+1)\xi} + e^{-i(2k+1)\xi}) \\ + \left(\cos^2 \frac{\xi}{2}\right)^K \left(\sin^2 \frac{\xi}{2}\right)^K f(\xi). \end{aligned} \tag{4.19}$$

Let us concentrate on the first two terms in (4.19). For large K, the coefficient of $e^{i(2k+1)\xi}$ tends to

$$2^{-4K+1} K \binom{2K}{K} \frac{(-1)^k}{2k+1} \binom{2K-1}{K+k} \underset{K\to\infty}{\sim} \frac{(-1)^k}{\pi(2k+1)},$$

which is exactly the Fourier coefficient of the characteristic function $\chi(\xi)=1$ for $|\xi|\leqq\pi/2$, 0 for $|\xi|\geqq\pi/2$,

$$\chi(\xi)=\frac{1}{2}+\sum_{k=0}^{\infty}(-1)^k\frac{1}{\pi(2k+1)}(e^{i(2k+1)\xi}+e^{-i(2k+1)\xi}).$$

This is, in fact, a perfectly legitimate choice for m_0: $m_0=\chi$ leads to $\hat{\phi}(\xi)=1$ for $|\xi|\leqq\pi, 0$ otherwise, or $\phi(x)=\sin\pi x/\pi x$. The corresponding wavelet basis is C^∞, satisfies (4.1) for arbitrarily large L, but has rather slow decay at ∞. Our ansatz (4.7) or (4.19) for m_0 can, therefore, be viewed as a truncation to finite length of χ, consistent with the restrictions (4.2), (4.3), and where an additional f has to be introduced to fit (1.9). Since for $K\to\infty$, χ itself already satisfies all the conditions (1.9), (4.2), (4.3), it seems reasonable to hope that for large K, a slight perturbation of χ might satisfy (1.9), (4.2), (4.3).

Based on this perturbation argument, we can look for a solution to (4.18) "close to" $x_m\equiv 0$. For $K=1,2,3,4$, and 5 we have (1°) determined the system (4.18) with the symbolic manipulation package MACSYMA, (2°) found a solution by Newton's method, starting from the initial point $x_m\equiv 0$, $m=0,\dots,K-1$. The resulting m_0 are tabulated in Table 2. For $K=5$ the coefficients are given with less precision than for $K\leqq 4$ because the roundoff error, even with double precision, was sufficient to perturb decimals beyond the 10th decimal. Note that Table 2 corrects a mistake in the first entry in the corresponding Table 8.1 in [25]. Graphs for the corresponding ϕ, ψ can be found in [25, Fig. 8.3].

Remarks.

(1) The functions ϕ and ψ corresponding to Table 2 are almost symmetric. For some of these examples, there exists a pair of biorthogonal bases very close to the orthonormal basis (their graphs are almost indistinguishable), which have, moreover, the advantage of corresponding to rational c_n (see [21]).

(2) The approach given above has the merit of giving a method for the construction of coiflets of any order L (modulo the solution of a system of $L/2$ quadratic equations in $L/2$ variables). It does not necessarily give the smoothest coiflet of order L, however! For small L, everything can be worked out more or less by hand, and we find some solutions different from the coiflets given above.

For $L=2$, the smoothest coiflet is found by substituting

$$f(\xi)=a\,e^{i\xi}+b\,e^{2i\xi},$$

rather than (4.8) into (4.7), leading to a less symmetric coiflet with support width 5; in this case support $\phi=[-1,4]$. The system of quadratic equations reduces to a single equation, so that everything can be solved explicitly. We find

$$a=(s-1)/2,\quad b=(-s+3)/2,\quad \text{with } s=\pm\sqrt{15}.$$

The choice $s=-\sqrt{15}$ gives the most regular coiflet of order 2. The corresponding ϕ is plotted in Fig. 5. This ϕ is continuously differentiable; using the methods of [11] we find that its derivative has Hölder exponent .191814... .

For $L=4$, we find, unlike the $L=2$ case, that the best regularity for ϕ is achieved by distributing its support as symmetrically as possible. This corresponds to choosing

$$f(\xi)=a\,e^{-i\xi}+b+c\,e^{i\xi}+d\,e^{2i\xi}.$$

The resulting set of equations reduces to two linear and two quadratic equations. All this can be reduced to one equation for a of degree 4, which has 2 real and 2 complex solutions. One of the real roots leads to a twice continuously differentiable ϕ,

TABLE 2

The coefficients for coiflets of order $2K$, $K = 1$ *or* 5. *Note: In this table, the coefficients are normalized so that their sum is* 1.

	n	$\frac{1}{2}c_n$
$K = 1$	−2	−.051429484095
	−1	.238929728471
	0	.602859456942
	1	.272140543058
	2	−.051429972847
	3	−.011070271529
$K = 2$	−4	.011587596739
	−3	−.029320137980
	−2	−.047639590310
	−1	.273021046535
	0	.574682393857
	1	.294867193696
	2	−.054085607092
	3	−.042026480461
	4	.016744410163
	5	.003967883613
	6	−.001289203356
	7	−.000509505399
$K = 3$	−6	−.002682418671
	−5	.005503126709
	−4	.016583560479
	−3	−.046507764479
	−2	−.043220763560
	−1	.286503335274
	0	.561285256870
	1	.302983571773
	2	−.050770140755
	3	−.058196250762
	4	.024434094321
	5	.011229240962
	6	−.006369601011
	7	−.001820458916
	8	.000790205101
	9	.000329665174
	10	−.000050192775
	11	−.000024465734
$K = 4$	−8	.000630961046
	−7	−.001152224852
	−6	−.005194524026
	−5	.011362459244
	−4	.018867235378
	−3	−.057464234429
	−2	−.039652648517
	−1	.293667390895
$K = 4$	0	.553126452562
	1	.307157326198
	2	−.047112738865
	3	−.068038127051
	4	.027813640153
	5	.017735837438
	6	−.010756318517
	7	−.004001012886
	8	.002652665946
	9	.000895594529
	10	−.000416500571
	11	−.000183829769
	12	.000044080354
	13	.000022082857
	14	−.000002304942
	15	−.000001262175
$K = 5$	−10	−.0001499638
	−9	.0002535612
	−8	.0015402457
	−7	−.0029411108
	−6	−.0071637819
	−5	.0165520664
	−4	.0199178043
	−3	−.0649972628
	−2	−.0368000736
	−1	.2980923235
	0	.5475054294
	1	.3097068490
	2	−.0438660508
	3	−.0746522389
	4	.0291958795
	5	.0231107770
	6	−.0139736879
	7	−.0064800900
	8	.0047830014
	9	.0017206547
	10	−.0011758222
	11	−.0004512270
	12	.0002137298
	13	.0000993776
	14	−.0000292321
	15	−.0000150720
	16	.0000026408
	17	.0000014593
	18	−.0000001184
	19	−.0000000673

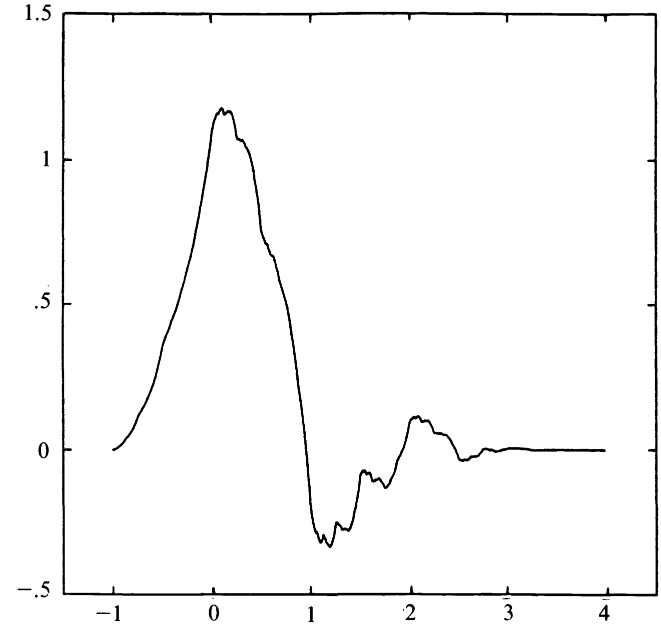


FIG. 5. *Plot of ϕ for the coiflet of order 2 with the highest regularity.*

corresponding to

$$\begin{array}{rlrl}
c_{-5}= & -.008089728693, & c_1= & .503931298301,\\
c_{-4}= & -.001473073456, & c_2= & .443259223184,\\
c_{-3}= & .027620978693, & c_3= & .010862015621,\\
c_{-2}= & .000661782050, & c_4= & -.136801026363,\\
c_{-1}= & -.029586627843, & c_5= & -.004737936078,\\
c_0= & .168333606358, & c_6= & .026019488227.
\end{array}$$

As in Table 2, these c_n are normalized so that their sum equals 1. We have plotted the corresponding ϕ in Fig. 6.

For $L=6$ explicit computation of all the solutions is more complicated, but still feasible. There exists no solution so that support $(\phi)=[-8,9]$. For the ansatz

$$f(\xi)=a\,e^{-i\xi}+b+c\,e^{i\xi}+d\,e^{2i\xi}+e\,e^{3i\xi}+f\,e^{4i\xi},$$

corresponding to support $(\phi)=[-7,10]$, there are two solutions. The most regular of these solutions is twice differentiable; it is given by

$$\begin{array}{rlrl}
c_{-7}= & -.000152916987, & c_2= & .269094527854,\\
c_{-6}= & .000315249697, & c_3= & .558133106629,\\
c_{-5}= & .001443474332, & c_4= & .322997271647,\\
c_{-4}= & -.001358589300, & c_5= & -.040303265359,\\
c_{-3}= & -.007915890196, & c_6= & -.069655118535,\\
c_{-2}= & .006194347829, & c_7= & .015323777973,\\
c_{-1}= & .025745731466, & c_8= & .013570199856,\\
c_0= & -.039961569717, & c_9= & -.002466300927,\\
c_1= & -.049807716931, & c_{10}= & -.001196319329.
\end{array}$$

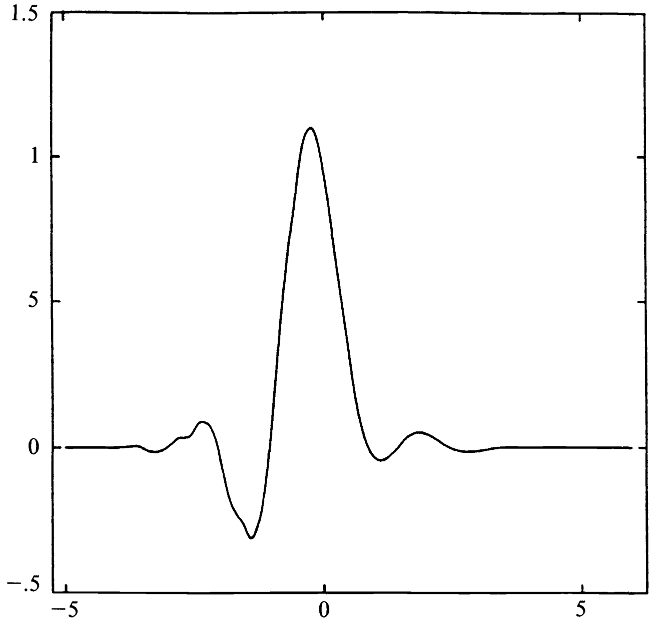


FIG. 6. *Plot of ϕ for the most regular coiflet of order* 4.

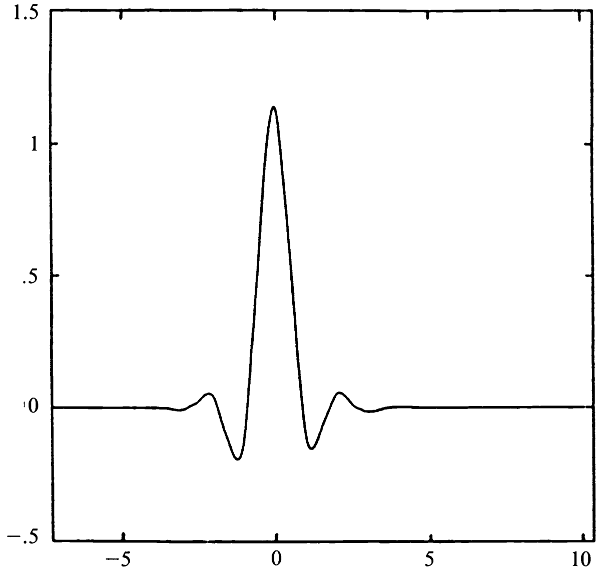


FIG. 7. *Plot of ϕ for the most regular coiflet of order* 6, *with support* $\phi = [-7, 11]$.

The function ϕ is plotted in Fig. 7. The coiflets used in [19] for $L = 2, 4, 6$ correspond to the scaling functions ϕ in Figs. 5, 6, and 7.

Acknowledgments. I would like to thank everyone who suggested the usefulness of the variations constructed in this paper, in particular, M. Basseville, G. Beylkin, A. Cohen, R. Coifman, and Y. Meyer.

REFERENCES

[1] J. Morlet, *Sampling theory and wave propagation*, in Issues on Acoustic Signal/Image Processing and Recognition, C. H. Chen, ed., NATO ASI, Springer-Verlag, New York, 1983.

[2] A. Grossmann and J. Morlet, *Decomposition of Hardy functions into square integrable wavelets of constant shape*, SIAM J. Math. Anal., 15 (1984), pp. 723-736.

[2a] P. Goupillaud, A. Grossmann, and J. Morlet, *Cycle-octave and related transforms in seismic signal analysis*, Geoexploration, 23 (1984), p. 85.

[3] A. Grossmann, J. Morlet, and T. Paul, *Transforms associated to square integrable group representations*, I, J. Math. Phys., 26 (1985), pp. 2473-2479; II, Ann. Inst. H. Poincaré, 45 (1986), pp. 293-309.

[4] I. Daubechies, A. Grossmann, and Y. Meyer, *Painless non-orthogonal expansions*, J. Math. Phys., 27 (1986), pp. 1271-1283.

[4a] ———, *The wavelet transform, time-frequency localization and signal analysis*, IEEE Trans. Inform. Theory, 34 (1988), pp 605-612.

[5] P. Auscher, *Ondelettes fractales et applications*, Ph.D. thesis, CEREMADE, University of Paris IX, 1989.

[6] S. Mallat, *Multiresolution approximation and wavelets*, Trans. Amer. Math. Soc., 315 (1989), pp. 69-88.

[7] Y. Meyer, *Ondelettes, function splines, et analyses graduées*, Lectures given at the Mathematics Department, University of Torino, 1986.

[8] G. Battle, *A block spin construction of ondelettes. Part* I: *Lemairé functions*, Comm. Math. Phys., 110 (1987), pp. 601-615.

[9] P. G. Lemarié, *Une nouvelle base d'ondelettes de $L^2(\mathbb{R}^n)$*, J. Math. Pures Appl., to appear.

[10] Y. Meyer, *Ondelettes, opérateurs et analyse non linéaire*, Hermann, Paris, 1990.

[11a] I. Daubechies and J. Lagarias, *Two-scale difference equations*, I. *Global regularity of solutions*, SIAM J. Math. Anal., 22 (1991), pp. 1388-1410.

[11b] ———, *Two-scale difference equations*, II. *Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031-1079.

[12] W. Lawton, *Tight frames of compactly supported wavelets*, J. Math. Phys., 31 (1990), pp. 1898-1901.

[13] A. Cohen, *Ondelettes, analyses multirésolution et filtres mirroir en quadrature*, Ann. Inst. Poincaré, Analyse non linéaire, 7 (1990), pp. 439-459.

[14] G. Deslauriers and S. Dubuc, *Interpolation dyadique*, in Fractals, dimensions non entières et applications, G. Cherbit, ed., Masson, Paris, 1987, pp. 44-56.

[15] I. Daubechies, *Orthonormal basis of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909-996.

[16] *Wavelets—time-frequency methods and phase space*, Proceedings of the December '87 Conference, Marseille, France, J. M. Combes, A. Grossmann, and Ph. Tchamitchian, eds., Springer, Berlin, 1989.

[17] G. Polya and G. Szegö, *Aufgaben und Lehrsätze aus der Analysis*, Vol. II, Springer, Berlin, 1971.

[18] G. Battle and P. Federbush, *Ondelettes and phase cell cluster expansions: a vindication*, Comm. Math. Phys., 109 (1987), pp. 417-419.

[19] G. Beylkin, R. Coifman, and V. Rokhlin, *Fast wavelet transforms and numerical algorithms*. I, Comm. Pure Appl. Math., 44 (1991), pp. 141-183.

[20] Y. Meyer, *Wavelets with compact support*, Zygmund lectures, University of Chicago, Chicago, IL, 1987.

[21] A. Cohen, I. Daubechies, and J. C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485-560.

[22] G. Battle, *Phase space localization theorem for ondelettes*, J. Math. Phys., 30 (1989), pp. 2195-2196.

[23] A. Cohen and J. P. Conze, *Régularité des bases d'ondelettes et mesures ergodiques*, Rev. Mat. Iberoamericana, to appear.

[24] H. Volkmer, *On the regularity of wavelets*, IEEE Trans. Inform. Theory, 38 (1992), pp. 872-876.

[25] I. Daubechies, *Ten lectures on wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[26] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics*, Addison-Wesley, Reading, MA, 1989.

SIAM J. MATH. ANAL.
Vol. 24, No. 2, pp. 520-527, March 1993

015

ORTHONORMAL BASES OF COMPACTLY SUPPORTED WAVELETS III. BETTER FREQUENCY RESOLUTION*

A. COHEN† AND INGRID DAUBECHIES‡

Abstract. Standard orthonormal bases of wavelets with dilation factor 2 use wavelets with one octave bandwidth. Orthonormal bases with $\frac{1}{2}$ octave or even smaller bandwidth wavelets are constructed. These wavelets are special cases of the "wavelet packet" construction by R. Coifman and Y. Meyer [Yale University, preprint, 1990].

Key words. wavelets, orthonormal bases, frequency resolution

AMS(MOS) subject classifications. 26A16, 26A18, 26A27, 39B12

1. Introduction. In the preceding paper [1] one of us showed how to construct variations to [2] in order to obtain orthonormal bases of compactly supported wavelets with various desirable properties. In this paper we show how to construct orthonormal bases with better frequency localization.

We shall continue to use here the notation and terminology given in [1, § 1]. When we need equation $(1.n)$ in [1], we shall refer to it as [1, $(1.n)$].

2. Statement of the problem. The Fourier transform $\hat{\psi}$ and ψ for an orthonormal wavelet basis of type [1, (1.1)] is, in most cases of practical interest, concentrated around a frequency band $a \leqq |\omega| \leqq 2a$. This "concentration" should be understood more or less loosely, depending on the example. For the Meyer basis [3], where $\hat{\psi}$ has compact support, we find support $(\hat{\psi}) \subset \{\omega;\ \pi - \varepsilon \leqq |\omega| \leqq 2\pi + 2\varepsilon\}$ for some $\varepsilon \in]0, \pi/3]$ (the "standard" choice is $\varepsilon = \pi/3$). For the functions constructed in [2], support$(\hat{\psi}) = \mathbb{R}$ (since support(ψ) is compact), but graphs of $\hat{\psi}$ show a reasonably good concentration around $\pi \leqq |\omega| \leqq 2\pi$ (see Fig. 7 in [2]).

It is the use of the dilation factor 2 in the definition [1, (1.1)] of the orthonormal basis that forces ψ to have a bandwidth of at least one octave. In many applications, especially those where a stationary high frequency component can be present (texture in images, music in acoustical signal analysis), it is desirable to have better frequency localization [4]. We present several approaches to the construction of such bases.

3. Noninteger dilation factors. Since the one-octave bandwidth is forced by the use of powers of 2, a natural way to obtain smaller bandwidth is to use a smaller dilation factor, e.g., $\frac{3}{2}$. There do indeed exist orthonormal wavelet bases for noninteger rational dilation factors α (for $\alpha = (k+1)/k$, $k \in \mathbb{N}$ this is an extension of Meyer's construction proposed by David [3]; for $\alpha = p/q$, a construction method is given in Auscher's Ph.D. thesis [5]). Unfortunately, these cannot correspond to a multiresolution analysis with compactly supported ϕ and ψ. For $\alpha = \frac{3}{2}$, e.g., $V_0 \subset V_{-1}$ implies the

* Received by the editors May 29, 1990; accepted for publication (in revised form) May 23, 1992.

† Centre de la Recherche de Mathematique de la Decision, Université Paris IX Dauphine, 75016 Paris, France.

‡ Mathematics Department, Rutgers University, New Brunswick, New Jersey 08903, and AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974. Part of this author's work was done while visiting the University of Paris IX, Paris-Dauphine.

existence of two sequences, $(c_{1,n})_{n\in\mathbb{Z}}$ and $(c_{2,n})_{n\in\mathbb{Z}}$, such that

$$\phi(x)=\sum_n c_{1,n}\phi\left(\frac{3}{2}x-n\right),$$
(1)
$$\phi(x-1)=\sum_n c_{2,n}\phi\left(\frac{3}{2}x-n\right).$$

If ϕ were compactly supported, then both sequences would be finite. But (1) can be rewritten as

$$\frac{3}{2}\hat{\phi}\left(\frac{3}{2}\xi\right)=\left(\sum_n c_{1,n}\,e^{in\xi}\right)\hat{\phi}(\xi),$$

$$\frac{3}{2}e^{i3\xi/2}\hat{\phi}\left(\frac{3}{2}\xi\right)=\left(\sum_n c_{2,n}\,e^{in\xi}\right)\hat{\phi}(\xi),$$

implying

$$\sum_n c_{1,n}\,e^{in\xi}=e^{i3\xi/2}\sum_n c_{2,n}\,e^{in\xi},$$

which is impossible for finite sequences. The same phenomenon occurs for any noninteger α.

In practical applications, it is desirable to have finite sequences c_n for the hierarchic decomposition + reconstruction scheme sketched in [1, § 1]. We can, of course, always use truncated versions of infinite sequences, but untruncated finite sequences are preferable. Therefore, noninteger α does not seem to be a good solution. (Note, however, that Kovačević and Vetterli [10] have constructed subband filtering schemes with FIR filters and rational noninteger dilation factors; these do not correspond to a multiresolution analysis with a single ϕ.)

4. Integer dilation factors larger than 2. Another candidate for wavelet bases with better frequency localization is given by constructions with a dilation factor *larger* than 2. In general, a multiresolution scheme with dilation factor N uses one scaling function ϕ, and $N-1$ different wavelets ψ^l, $l=1,\dots,N-1$ [5]. The function ϕ satisfies an equation of type [1, eqn. (1.4)], $\phi(x)=\sum_n c_n\phi(Nx-n)$, or $\hat{\phi}(\xi)=m_0(\xi/N)\hat{\phi}(\xi/N)$, with $m_0(\xi)=N^{-1}\sum_n c_n\,e^{in\xi}$. The different ψ^l can all be written as linear combinations of the $\phi(Nx-n)$. There exist, therefore, trigonometric polynomials m_l so that $\hat{\psi}^l(\xi)=m_l(\xi/N)\hat{\phi}(\xi/N)$ (we assume compact support for ϕ and the ψ^l, so that m_0 and the m_l are indeed polynomials and not infinite series). Orthonormality of the different subspaces in the multiresolution analysis then implies that the $N\times N$ matrix $\mathbb{M}(\xi)$ with entries

$$\mathbb{M}_{lk}(\xi)=m_{l-1}\left(\xi+\frac{2\pi}{N}(k-1)\right)$$

is unitary for all ξ [5]. For $N=2$, this reduces to the standard requirements $m_1(\xi)=e^{i\xi}\overline{m_0(\xi+\pi)}\lambda(\xi)$, with $|\lambda(\xi)|=1$, and $\lambda(\xi+\pi)=\lambda(\xi)$, and $|m_0(\xi)|^2+|m_0(\xi+\pi)|^2=1$. The second condition is [1, eqn. (1.9)] again. If both m_0 and m_1 are trigonometric polynomials, then the only possible choices for $\lambda(\xi)$ are $\lambda(\xi)=\lambda\,e^{ik\xi}$, with $\lambda\in\mathbb{C}$, $|\lambda|=1$ and $k\in\mathbb{Z}$. The simplest choice $\lambda(\xi)\equiv 1$ corresponds to [1, eqn. (1.6)].

For general N, the step of one resolution space V_j to the coarser space V_{j+1} corresponds to a jump of $\log_2 N$ octaves in frequency, since there is a dilation factor N between the two resolutions, and every factor 2 corresponds to one octave. Each of the $N-1$ wavelets ψ^l corresponds therefore to, on average, a bandwidth of

$(N-1)^{-1}\log_2 N$. For $N=4$, for instance, we would have 3 wavelets for 2 octaves, corresponding to an average $\frac{2}{3}$-octave per wavelet, which is better than the 1-octave bandwidth attained by bases with $N=2$.

This naive computation of the average bandwidth of the ψ^l is deceptive, however, as illustrated by the following example. Given an orthonormal wavelet basis with dilation factor 2, the following easy trick generates an orthonormal wavelet basis with dilation factor 4. Let us, for this paragraph only, use a tilde to distinguish the wavelets, spaces in the multiresolution analysis, etc.... for the $N=4$ construction from their counterparts in the $N=2$ case. Define

$$\tilde{m}_0(\xi)=m_0(\xi)m_0(2\xi),$$
$$\tilde{m}_1(\xi)=m_0(\xi)m_1(2\xi),$$
$$\tilde{m}_2(\xi)=m_1(\xi)m_1(2\xi),$$
$$\tilde{m}_3(\xi)=m_1(\xi)m_0(2\xi),$$

where $m_1(\xi)=e^{i\xi}\overline{m_0(\xi+\pi)}$, and $m_0(\xi)$ is the trigonometric polynomial associated with the given orthonormal wavelet basis with $N=2$. Because m_0 satisfies [1, eqn. (1.9)], we easily check that the 4×4 matrix $\tilde{\mathbb{M}}(\xi)$ is indeed unitary for all ξ. The corresponding function $\tilde{\phi}$ is given by

$$\hat{\tilde{\phi}}(\xi)=\prod_{j=1}^{\infty}\tilde{m}_0(4^{-j}\xi)=\prod_{j=1}^{\infty}[m_0(2^{-2j}\xi)m_0(2^{-2j+1}\xi)]=\prod_{j=1}^{\infty}m_0(2^{-j}\xi)=\hat{\phi}(\xi),$$

so that $\tilde{\phi}\equiv\phi$. The scaling spaces $\tilde{V}_j$, spanned by the $\tilde{\phi}(4^{-j}x-k)$, are, therefore, the subsequence of scaling spaces V_m with even index, $\tilde{V}_j=V_{2j}$. The wavelets $\tilde{\psi}^l$, $l=1,2,3$ are given by

$$\hat{\tilde{\psi}}^l(\xi)=\hat{\tilde{m}}_l\left(\frac{\xi}{4}\right)\hat{\phi}\left(\frac{\xi}{4}\right).$$

In particular,

$$\hat{\tilde{\psi}}^1(\xi)=m_1\left(\frac{\xi}{2}\right)m_0\left(\frac{\xi}{4}\right)\hat{\phi}\left(\frac{\xi}{4}\right)=m_1\left(\frac{\xi}{2}\right)\hat{\phi}\left(\frac{\xi}{2}\right)=\hat{\psi}(\xi),$$

or $\tilde{\psi}^1\equiv\psi$. Since the function ψ is essentially a 1-octave bandwidth function, this shows that the computation above of the average bandwidth of $\frac{2}{3}$-octave for the $\tilde{\psi}^l$ was indeed naive and misleading. The other two wavelets, $\tilde{\psi}^2$, $\tilde{\psi}^3$, are functions not existing in the $N=2$ case, which have in fact better frequency localization. They generate spaces $\tilde{W}_j^2$, $\tilde{W}_j^3$, which together with $\tilde{W}_j^1=W_{2j}$ complement $\tilde{V}_j=V_{2j}$ to constitute $\tilde{V}_{j-1}=V_{2j-2}$,

$$V_{2j-2}=V_{2j}\oplus W_{2j}\oplus\tilde{W}_j^2\oplus\tilde{W}_j^3.$$

Since $V_{2j}\oplus W_{2j}=V_{2j-1}$, it follows that $\tilde{W}_j^2\oplus\tilde{W}_j^3=W_{2j-1}$. This construction of an $N=4$ multiresolution analysis from an $N=2$ case corresponds therefore to a splitting of all the odd-indexed W_{2j-1} into two spaces, each of approximately $\frac{1}{2}$-octave bandwidth functions, while the even-index W_{2j} remain untouched, and are still generated by 1-octave bandwidth functions.

In the next section we show how to do better than this construction by splitting *every* W_j (as opposed to only the odd-indexed ones) into two parts.

5. Wavelet bases with $\frac{1}{2}$-octave bandwidth wavelets. At the end of the previous section, we had split W_1 into two $\frac{1}{2}$-octave spaces, $\tilde{W}_1^2\oplus\tilde{W}_1^3$. Dyadic dilations of these spaces are then $\frac{1}{2}$-octave components for any W_m, and these spaces can be used to decompose all of $L^2(\mathbb{R})$. This decomposition can also be derived directly. The key idea is to start from a conventional multiresolution analysis with dilation factor 2, and to split every space W_j into two $\frac{1}{2}$-octave bandwith subspaces. They key to this splitting is [1, eqn. (1.8)] and the computation leading to [1, eqn. (1.11)]. As shown in [1, § 1], condition [1, (1.8)] ensures that [1, (1.4)] and [1, (1.6)] define an orthonormal basis transform in V_{-1}, from $\{\sqrt{2}\phi(2x-n);\ n\in\mathbb{Z}\}$ to $\{\phi(x-n), \psi(x-n);\ n\in\mathbb{Z}\}$. For the argument to work, it is not essential that the functions appearing in the two sides of [1, (1.4)] are dilated versions of each other. The same argument would work just as well for *any* function f such that the $f(x-k)$ constitute an orthonormal set; with the definitions

$$g_1(x)=2^{-1/2}\sum_n c_n f(x-n),$$

$$g_2(x)=2^{-1/2}\sum_n (-1)^n c_{-n+1} f(x-n),$$

the functions $(g_1(x-2n), g_2(x-2m))_{n,m\in\mathbb{Z}}$ are another orthonormal basis for $\overline{\text{Span}\,(f(x-k))}$.

Given α_n satisfying [1, (1.8)], i.e., $\sum_n \alpha_n\alpha_{n+2k}=2\delta_{k0}$, we can therefore define, for any basic wavelet ψ, the functions $\tilde{\psi}_1$, $\tilde{\psi}_2$ by

$$\tilde{\psi}_1(x)=2^{-1/2}\sum_n \alpha_n\psi(x-n)$$

$$\tilde{\psi}_2(x)=2^{-1/2}\sum_n (-1)^n\alpha_{-n+1}\psi(x-n); \tag{2}$$

the functions $\tilde{\psi}_1(x-2n)$, $\tilde{\psi}_2(x-2n)$ then constitute an orthogonal basis for W_0; consequently $(2^{-j/2}\tilde{\psi}_1(2^{-j}x-2n), 2^{-j/2}\tilde{\psi}_2(2^{-j}x-2n);\ j, n\in\mathbb{Z})$ is an orthonormal basis for $L^2(\mathbb{R})$. The Fourier transforms of $\tilde{\psi}_1$, $\tilde{\psi}_2$ will be concentrated on respectively the higher and lower half of the bandwidth of ψ. The bandwidth of $\tilde{\psi}_1$ is, therefore, approximately $\log_2\frac{4}{3}=2-\log_2 3$ octaves, that of $\tilde{\psi}_2$ approximately $\log_2\frac{3}{2}=\log_2 3-1$ octaves. The supportwidth of $\tilde{\psi}_1$, $\tilde{\psi}_2$ will of course be larger than that of ψ.

In Fig. 1 we show two different constructions of $\tilde{\psi}_1$, $\tilde{\psi}_2$, built from two different functions ψ. In each case we have chosen for α_n the same coefficients c_n as used for the definition of ψ (via [1, (1.4)], [1, (1.6)]). We may, of course, also choose different α_n. The support width of $\tilde{\psi}_1$, $\tilde{\psi}_2$ is twice as large as that of ψ. In general, we have

$$|\text{support}(\tilde{\psi}_1)|=|\text{support}\,(\tilde{\psi}_2)|=(\#\text{ of nonzero }\alpha_n)-1$$
$$+(\#\text{ of nonzero }c_n)-1.$$

For the choice $\alpha_n=c_n$ made in Fig. 1, we have, therefore,

$$|\text{support}\,(\tilde{\psi}_1)|=|\text{support}\,(\tilde{\psi}_2)|=2|\text{support}\,(\psi)|.$$

Note that with the choice $\alpha_n=c_n$ the two wavelets constructed here are (up to dilation) exactly the two "new" wavelets constructed in the previous section.

The plots of $|\hat{\tilde{\psi}}_1|$, $|\hat{\tilde{\psi}}_2|$ in Fig. 1 have "side-lobes" that make the splitting less than perfect. Such side-lobes are unavoidable. We have

$$\hat{\tilde{\psi}}_1(\xi)=a(\xi)\hat{\psi}(\xi), \qquad \hat{\tilde{\psi}}_2(\xi)=e^{i\xi}\overline{a(\xi+\pi)}\hat{\psi}(\xi),$$

with $a(\xi)=2^{-1/2}\sum_n \alpha_n e^{in\xi}$. Condition [1, (1.8)] implies $|a(\xi)|^2+|a(\xi+\pi)|^2=2$. For small $|\xi|\leqq\xi_0$ we have $|\hat{\tilde{\psi}}_1|\simeq\sqrt{2}|\hat{\psi}|$, or $|a(\xi)|\simeq\sqrt{2}$. Consequently, $|a(\xi+\pi)|\simeq 0$ for $|\xi|\leqq\xi_0$. Since a is periodic, this implies $|a(\xi+\pi)|\simeq 0$ for $|\xi-2\pi|\leqq\xi_0$, or $|a(\xi)|\simeq\sqrt{2}$, hence $|\hat{\tilde{\psi}}_1|\simeq\sqrt{2}|\hat{\psi}|$ in this region also. This causes the "side-lobes" on the figure. The only way to reduce them is to start from ψ with very concentrated Fourier transform

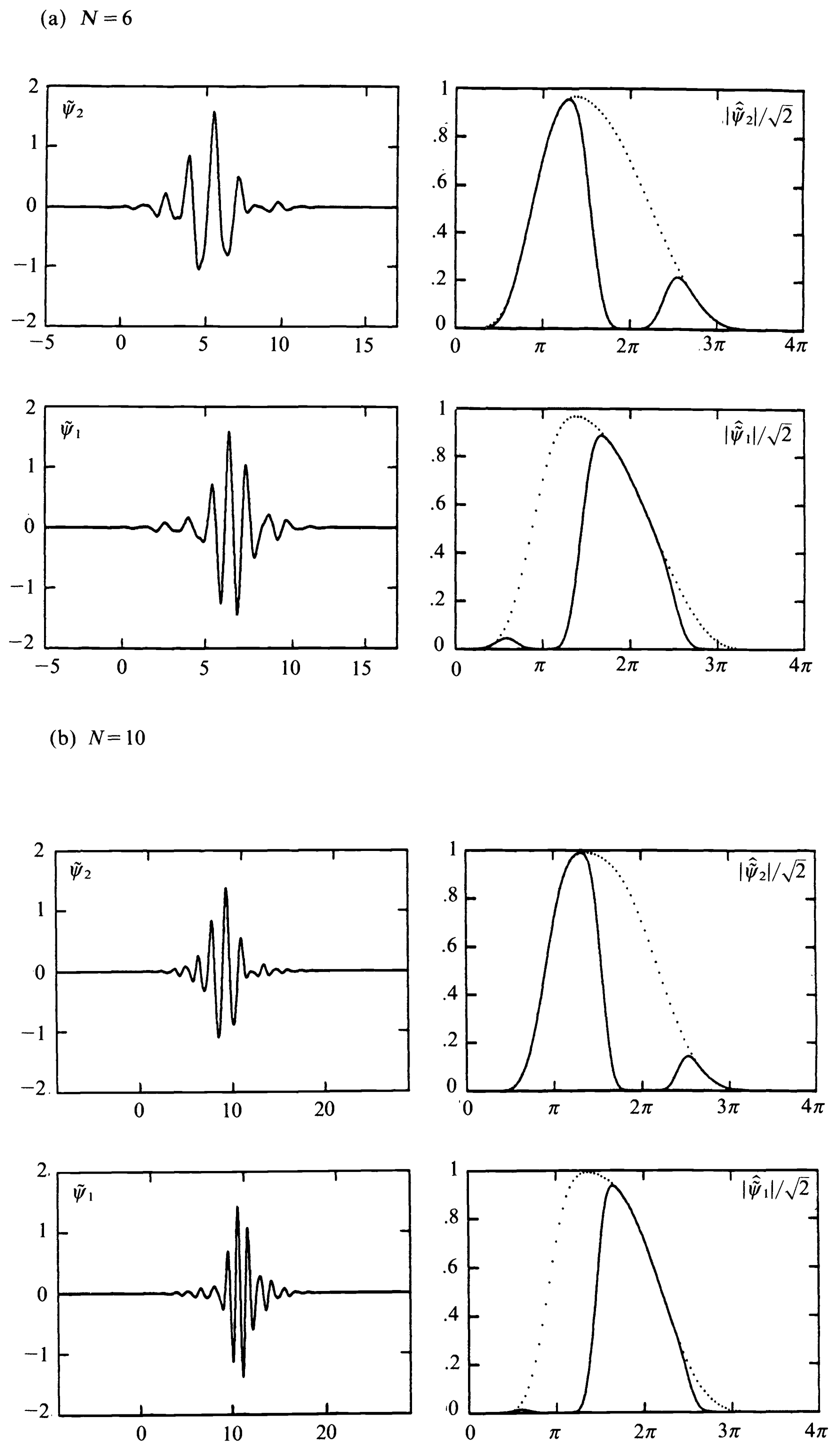


FIG. 1. *Graphs of the "$\frac{1}{2}$-octave bandwidth" functions* $\tilde\psi_1$, $\tilde\psi_2$ *and the absolute values of their Fourier transforms* $|\hat{\tilde\psi}_1|$, $|\hat{\tilde\psi}_2|$. *The dotted line in each case is the graph of* $|\hat\psi|$. *We have* $2|\hat\psi|^2 = |\hat{\tilde\psi}_1|^2 + |\hat{\tilde\psi}_2|^2$. (*This is clearly illustrated by the graphs. Note that we have plotted* $|\hat{\tilde\psi}_1|/\sqrt{2}$ *and* $|\hat{\tilde\psi}_2|/\sqrt{2}$ *in order to make the comparison with* $|\hat\psi|$.) *In both cases the* α_n (*see* (2)) *are the same as the coefficients* c_n *from which* ψ *is defined*; *the function* ψ *in these two examples corresponds to* (a) *the case* $N=6$ *in* [1, § 2] *and* (b) *the case* $N=10$ *in* [1, § 2].

$|\hat{\psi}|$. This is why the effect becomes less pronounced as N increases. (The Fourier transform $|\hat{\psi}|$ becomes more concentrated as N increases; see Fig. 7 in [2].)

Remarks.

(1) In case even better frequency localization is desired, this splitting trick can be repeated: we can replace ψ by $2j$ functions, each corresponding to a 2^{-j}-octave bandwidth, and which have to be translated by integer multiples of 2^j,

$$\overline{\text{Span}\,\{\psi(x-n)\}}=\overline{\text{Span}\,\{\tilde{\psi}_l(x-2^jm);\ l=1,\dots,2^j,\ m\in\mathbb{Z}\}}.$$

At every splitting, the "side-lobe effect" takes its toll, however, marring the frequency localization; see [8].

(2) The functions $\tilde{\psi}_1$, $\tilde{\psi}_2$ constructed are special cases of the "wavelet packets" discovered by Coifman and Meyer [6], which in one framework encompass many different choices of orthonormal bases, of which the wavelet is one extreme example. Another extreme within the same framework is a basis closer in spirit to the windowed Fourier transform; infinitely many intermediate choices are possible; see [9] for applications.

6. Multidimensional "splitting." The "splitting trick" also allows selective splitting in higher-dimensional multiresolution analysis. For the sake of simplicity we restrict ourselves to two dimensions. A standard way of generating a two-dimensional wavelet basis from a one-dimensional multiresolution analysis is to define [7]

$$\begin{aligned}\Phi(x,y)&=\phi(x)\phi(y),\\ \Psi^1(x,y)&=\psi(x)\phi(y),\\ \Psi^2(x,y)&=\phi(x)\psi(y),\\ \Psi^3(x,y)&=\psi(x)\psi(y).\end{aligned}\tag{3}$$

The orthonormal basis of wavelets is then given by the $\Psi^l_{j,k}$, with $l=1,2,3$, $j\in\mathbb{Z}$ and $k=(k_1,k_2)\in\mathbb{Z}^2$ defined by

$$\Psi^l_{j,k}(x,y)=2^{-j}\Psi^l(2^{-j}x-k_1,2^{-j}y-k_2).$$

A good way to visualize what this construction means is to look at it in the Fourier plane for the special choice $\hat{\phi}(\xi)=(2\pi)^{-1/2}$ if $|\xi|\leqq\pi$, 0 otherwise, and the corresponding $\hat{\psi}(\xi)=(2\pi)^{-1/2}$ if $\pi<|\xi|\leqq 2\pi$, 0 otherwise. We easily check that for this choice the space V_j is exactly $L^2([-2^{-j}\pi,2^{-j}\pi])$, and $W_j=L^2([-2^{-j+1}\pi,-2^{-j}\pi]\cup[2^{-j}\pi,2^{-j+1}\pi])$, so that the whole ladder of one-dimensional multiresolution spaces can be simply represented by the supports of the corresponding functions. Of course, these particular ϕ and ψ decay too slowly to be useful in practice, but the visualization they lead to in Fourier space is still "morally true" for other, more useful choices, even though the splitting is not as clean cut. The two-dimensional multiresolution spaces $\mathbf{V}_j$, and their complement spaces $\mathbf{W}^l_j$, generated by (3) can also be written

$$\begin{aligned}\mathbf{V}_j&=V_j\otimes V_j,\\ \mathbf{W}^1_j&=W_j\otimes V_j,\\ \mathbf{W}^2_j&=V_j\otimes W_j,\\ \mathbf{W}^3_j&=W_j\otimes W_j.\end{aligned}$$

For the special choices of ϕ, ψ described above, these two-dimensional spaces are again L^2-spaces of particular domains carved out in the Fourier plane. For instance, $\mathbf{V}_j = L^2([-2^{-j}\pi, 2^{-j}\pi]) \otimes L^2([-2^{-j}\pi, 2^{-j}\pi]) = L^2([-2^{-j}\pi, 2^{-j}\pi] \times [-2^{-j}\pi, 2^{-j}\pi])$; in particular, $\mathbf{V}_0 = L^2([-\pi, \pi]^2)$. For every j, the sum $\mathbf{W}_j^1 \oplus \mathbf{W}_j^2 \oplus \mathbf{W}_j^3$ is the orthogonal complement in $\mathbf{V}_{j-1}$ of $\mathbf{V}_j$, which corresponds to the annulus $[-2^{-j+1}\pi, 2^{-j+1}\pi]^2 \backslash [-2^{-j}\pi, 2^{-j}\pi]^2$. All this is visualized in Fig. 2(a): the small central square corresponds to (say) $\mathbf{V}_0$; adding to it the two vertical rectangles corresponding to $\mathbf{W}_0^1$, the two horizontal rectangles for $\mathbf{W}_0^2$, and the four corner squares for $\mathbf{W}_0^3$ lead to the bigger square representing $\mathbf{V}_{-1}$. The structure then repeats in the next annulus, to constitute $\mathbf{V}_{-2}$. The angular resolution in the Fourier plane of this scheme is not very good, as shown by the figure. Figure 2(b) shows what the same two-dimensional construction looks like when we start from a one-dimensional multiresolution analysis with dilation factor 4, as given in § 4. In this case the one-dimensional scheme has already three wavelets,

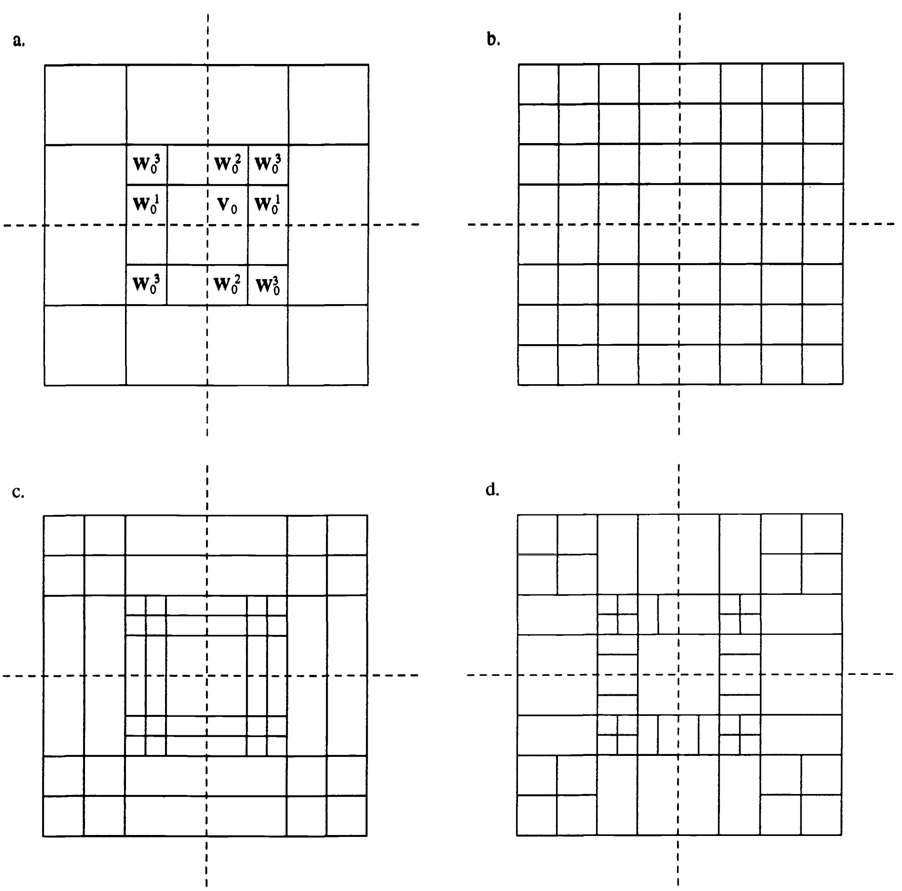


FIG. 2. *Visualization of the localization in the Fourier plane achieved by various two-dimensional multiresolution schemes*:

(a) *the standard product scheme starting from a one-dimensional analysis with dilation factor* 2,

(b) *product scheme from a one-dimensional analysis with dilation factor* 4, *derived from a multiresolution analysis with factor* 2 *as in* § 4,

(c) *product scheme from a one-dimensional analysis with two* $\frac{1}{2}$*-octave bandwidth wavelets rather than one* 1*-octave bandwidth wavelet*,

(d) *a nonproduct scheme obtained by the "splitting trick."*

so that the two-dimensional product scheme ends up with $2\times 3+3^2=15$ wavelets. Figure 2(b) represents one step (with dilation 4) in the multiresolution scale, as compared to two steps (with dilation factor 2), i.e., two successive annuli in Fig. 2(a). The central part of the two pictures is identical; the only difference between the two pictures is that the outer annulus of Fig. 2(a) is split into many pieces to give Fig. 2(b), while the inner annulus is untouched. This corresponds to the "splitting of one level out of two" shown at the end of § 4. The result is good angular resolution for some wavelets (corresponding to the outer layer in Fig. 2(b)), bad for others (corresponding to the most central rectangles in Fig. 2(b)).

Figure 2(c) shows the same picture again, with two steps in a multiresolution analysis with dilation factor 2, but for a product structure of type (3) starting from the two $\frac{1}{2}$-octave bandwidth wavelets constructed in § 5, rather than the 1-octave bandwidth wavelet ψ. The scaling function Φ is the same, but there are now $2\times 2+2^2=8$ wavelets (as opposed to 3 for Fig. 2(a), and 15 for Fig. 2(b)). Figure 2(c) can be obtained from Fig. 2(a) by splitting every annulus (inner as well as outer) into halves by cuts in both horizontal and vertical directions. This improves the angular resolution on the squares in the corners (corresponding to the $\mathbf{W}_j^3$ of Fig. 2(a), but does nothing for the angular resolution of the rectangles (corresponding to $\mathbf{W}_j^2$ or $\mathbf{W}_j^1$ in Fig. 2(a)), which were split better in the outer annulus of Fig. 2(b). The best angular resolution can be obtained by giving up a product structure analogous to (3) and just carving up every one of the $\mathbf{W}_j^l$ spaces of Fig. 2(a) vertically and/or horizontally, by applying the "splitting trick" in x and/or y, until the desired resolution is achieved. An example is given in Fig. 2(d). This still corresponds to an orthonormal basis, and to a fast algorithm for decomposing and reconstructing functions, as described in [1, § 1], although the organization is somewhat more complex. If even better angular resolution is required, then we can repeat the splitting trick as many times as necessary.

Acknowledgment. I. Daubechies would like to thank Professor Y. Meyer and the University of Paris IX for inviting her and for their hospitality.

REFERENCES

[1] I. DAUBECHIES, *Orthonormal bases of compactly supported wavelets* II. *Variations on a theme*, SIAM J. Math. Anal., this issue (1993), pp. 499–519.

[2] ———, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[3] Y. MEYER, *Principe d'incertitude, bases hilbertiennes et algèbres d'opérateurs.* Sém. Bourbaki, 662, 1985–1986.

[4] C. DORIZE AND K. GRAM-HANSEN, *Wavelet decomposition in the field of time-frequency analysis*, Conference on Wavelets, Marseille, France, June 1989.

[5] P. AUSCHER, *Ondelettes fractales et applications*, Ph.D. thesis, CEREMADE, University of Paris IX, Paris, France, 1989.

[6] R. COIFMAN AND MEYER, *Orthonormal Wavelet Packet Bases*, Yale University, New Haven, CT, preprint, 1990.

[7] S. MALLAT, *A theory for multiresolution signal decomposition: the wavelet representation*, IEEE Trans. on Pattern Analysis and Machine Intelligence, 11 (1989), pp. 674–693.

[8] R. COIFMAN, Y. MEYER, AND V. WICKERHAUSER, *Size properties of wavelet packets*, in Wavelets and their Applications, M. B. Ruskai et al., eds., Jones and Bartlett, Boston, MA, 1992, pp. 453–470.

[9] R. COIFMAN AND V. WICKERHAUSER, *Entropy-based algorithms for best basis selection*, IEEE Trans. Inform. Theory, 38 (1992), pp. 713–718.

[10] K. KOVAČEVIĆ AND M. VETTERLI, *Perfect reconstruction filter banks with rational sampling rates*, Columbia Univ., New York, preprint, 1991; IEEE Trans. Signal Processing, submitted.

SIAM J. MATH. ANAL.
Vol. 24, No. 2, pp. 528–556, March 1993

MARKOV–BERNSTEIN AND NIKOLSKII INEQUALITIES, AND CHRISTOFFEL FUNCTIONS FOR EXPONENTIAL WEIGHTS ON (–1,1)*

D. S. LUBINSKY[†] AND E. B. SAFF[‡]

Abstract. Exponential weights $w := e^{-Q}$ are considered, where $Q : (-1,1) \to \mathbf{R}$ is even, convex, and sufficiently smooth. For example, the results may be applied to

$$w(x) := (1-x^2)^{\alpha}, \qquad \alpha > 0,$$
$$w(x) := \exp(-(1-x^2)^{-\alpha}), \qquad \alpha > 0, \quad \text{or}$$
$$w(x) := \exp(-\exp_k(1-x^2)^{-\alpha}), \quad \alpha > 0, \quad k \geq 1,$$

where $\exp_k = \exp(\exp(\cdots))$ denotes the kth iterated exponential.

Weighted Markov and Bernstein inequalities such as

$$\|P'w\|_{L_\infty[-1,1]} \leq CQ'(a_{2n})\|Pw\|_{L_\infty[-1,1]},$$

and

$$|P'w|(x) \leq \frac{Cn}{\sqrt{1-|x|/a_n}}\|Pw\|_{L_\infty[-1,1]}, \qquad |x| < a_n,$$

are established for polynomials P of degree at most n. Here a_n is the nth Mhaskar–Rahmanov–Saff number for Q. For the special weights listed above, a more explicit form is given to $Q'(a_{2n})$. Estimates are deduced for Christoffel functions such as

$$\sup_{x\in[-1,1]} \lambda_n^{-1}(w^2,x)w(x) \leq CQ'(a_{2n}),$$

and also Nikolskii inequalities.

Key words. Markov–Bernstein inequalities, Nikolskii inequalities, non-Szegő weights, Christoffel functions, orthogonal polynomials

AMS(MOS) subject classifications. primary 41A17, 42C05; secondary 41A10

1. Introduction and statement of results. Throughout, $\mathcal{P}_n$ denotes the class of real polynomials of degree at most n, and $\|\cdot\|_{L_p(\mathcal{S})}$ denotes the L_p norm over any measurable $\mathcal{S} \subset \mathbf{R}$ $(0 < p \leq \infty)$. Furthermore, C, C_1, $C_2, \ldots,$ denote positive constants independent of n, $P \in \mathcal{P}_n$, and $x \in \mathbf{R}$, which are not necessarily the same in different occurrences.

The classical inequalities of Markov and Bernstein are, respectively,

$$\|P'\|_{L_\infty[-1,1]} \leq n^2\|P\|_{L_\infty[-1,1]} \tag{1.1}$$

and

$$|P'(x)| \leq n(1-x^2)^{-1/2}\|P\|_{L_\infty[-1,1]} \tag{1.2}$$

for $P \in \mathcal{P}_n$ and $|x| < 1$. The interest in these inequalities lies in their application to rates of approximation by polynomials, to discretisation procedures, to approximation processes, and so on.

* Received by the editors August 26, 1991; accepted for publication (in revised form) June 8, 1992.

[†] Department of Mathematics, University of Witwatersrand, P. O. Wits 2050, Republic of South Africa. This author's work was begun while visiting E. B. Saff at the University of South Florida in 1989.

[‡] Institute for Constructive Mathematics, Department of Mathematics, University of South Florida, Tampa, Florida 33620-5700. This author's research was supported in part by National Science Foundation grant DMS-881-4026.

Naturally, such important inequalities have been generalized to treat a host of situations, such as in L_p spaces, and with weights inserted. We cannot hope here to review the history and the contributions of the many authors. A fairly typical example of those in the literature is [1, Thm. 8.4.7, p. 107]:

$$\|P'(x)\sqrt{1-x^2}w(x)\|_{L_p[-1,1]} \leq Cn\|Pw\|_{L_p[-1,1]}, \tag{1.3}$$

$P \in \mathcal{P}_n$, $n \geq 1$. Here $0 < p \leq \infty$, and $w(x)$ is either a Jacobi weight $(1-x)^\alpha(1+x)^\beta$, where α, $\beta > -1$, or something similar. See [15, Thm. 19, p. 164], [1], [2], [3] for further discussion and references.

The aim of this paper is to treat weights $w(x)$ that may decay more rapidly at ± 1 than Jacobi weights; for example,

$$w(x) := W_{0,\alpha}(x) := \exp(-(1-x^2)^{-\alpha}), \qquad \alpha > 0, \tag{1.4}$$

or

$$w(x) := W_{k,\alpha}(x) := \exp(-\exp_k[(1-x^2)^{-\alpha}]), \quad \alpha > 0, \quad k \geq 1, \tag{1.5}$$

where

$$\exp_k := \exp(\exp(\exp\cdots)) \quad (k \text{ times})$$

denotes the kth iterated exponential. However, our results apply equally well to the classical ultraspherical weight

$$w(x) := (1-x^2)^\alpha, \qquad \alpha > 0.$$

For $\alpha = \frac{1}{2}$, the weight $W_{0,\alpha}$ of (1.4) is similar to a Pollaczek weight, and its orthogonal polynomials were considered in [15, pp. 82–83]. Asymptotics, and spacing of zeros of the orthogonal polynomials for $W_{k,\alpha}$, have been considered in [2], [12].

To the best of our knowledge, the Markov–Bernstein inequalities in this paper are new for the weights $W_{k,\alpha}$ for all the range of parameters. To those with an interest in orthogonal polynomials, it is noteworthy that Szegő's condition

$$\int_{-1}^{1} \frac{\log w(x)}{\sqrt{1-x^2}}dx > -\infty \tag{1.6}$$

is violated by $W_{0,\alpha}$ of (1.4) if $\alpha \geq \frac{1}{2}$ and by all the weights $W_{k,\alpha}$ of (1.5). Since the Markov–Bernstein inequalities have various applications to orthogonal polynomials associated with non-Szegő weights, such as estimates for Christoffel functions, they are of particular interest.

In fact, the results of this paper bear a close resemblance to results for exponential weights on the real line [4], [5], [6], [9], [10], [17], and more specifically to Erdös weights $W^2 = e^{-2Q}$ on $\mathbf{R}$: These have the property that Q grows faster than any polynomial at infinity.

In the analysis of those weights, and the ones treated in this paper, the Mhaskar–Rahmanov–Saff number plays an important role. Let us suppose that $w = e^{-Q}$, where $Q : (-1,1) \to \mathbf{R}$ is even, and differentiable in (0,1). Suppose, furthermore, that $tQ'(t)$ is positive and increasing in (0,1) with limits zero and infinity at zero and 1, respectively, and

$$\int_0^1 \frac{tQ'(t)}{\sqrt{1-t^2}}dt = \infty. \tag{1.7}$$

Then the uth Mhaskar–Rahmanov–Saff number, $a_u = a_u(Q)$, is defined to be the root of

$$(1.8) \qquad u = \frac{2}{\pi}\int_0^1 \frac{a_u t Q'(a_u t)}{\sqrt{1-t^2}}dt, \qquad u > 0.$$

The importance of a_u lies in the identity (cf. [13], [14])

$$(1.9) \qquad \|Pw\|_{L_\infty[-1,1]} = \|Pw\|_{L_\infty[-a_n,a_n]}, \quad P \in \mathcal{P}_n, \quad n \geq 1,$$

which we refer to as the *Mhaskar–Saff identity.* Of course $a_n \to 1$ as $n \to \infty$. As an illustration of its rate of approach, we note that for $w = W_{0,\alpha}$,

$$(1.10) \qquad 1 - a_n \sim n^{-1/(\alpha+1/2)}, \qquad n \to \infty,$$

and for $w = W_{k,\alpha}$,

$$(1.11) \qquad 1 - a_n \sim (\log_k n)^{-1/\alpha}, \qquad n \to \infty,$$

where $\log_k = \log(\log(\log\cdots))$ (k times) denotes the kth iterated logarithm. Furthermore, we are using $\sim$ in the sense of [15]: if $\{c_n\}_{n=1}^\infty$ and $\{d_n\}_{n=1}^\infty$ are real positive sequences, then

$$c_n \sim d_n, \qquad n \to \infty,$$

means that for n large enough,

$$C_1 \leq \frac{c_n}{d_n} \leq C_2.$$

We are now ready to define our class of weights.

DEFINITION 1.1. Let $w := e^{-Q}$, where

(i) Q is even and continuously differentiable in (-1,1), while Q'' is continuous in (0,1);

(ii) $Q' \geq 0$ and $Q'' \geq 0$ in (0,1);

(iii) $\int_0^1 (tQ'(t))/(\sqrt{1-t^2})dt = \infty$;

(iv) Let

$$(1.12) \qquad T(x) := 1 + \frac{xQ''(x)}{Q'(x)}, \qquad x \in (0,1).$$

We assume that

$$(1.13) \qquad \begin{array}{ll} \text{(a)} & T \text{ is increasing in } (0,1); \\ \text{(b)} & T(0+) > 1; \\ \text{(c)} & T(x) = O(Q'(x)),\ x \to 1^-. \end{array}$$

Under the above conditions, we write $w \in \mathcal{W}$.

We remark that (1.13) is a rather weak regularity condition, while (iii) is required for the existence of a_n. Further, we note that most of our results really only require the above hypotheses to be satisfied for x near 1. However, for simplicity, we shall not pursue this point. In any event, $W_{k,\alpha} \in \mathcal{W}$, $k \geq 0$, $\alpha > 0$.

Following are our Markov–Bernstein inequalities for $P'w$.

THEOREM 1.2. *Let $w \in \mathcal{W}$.*

(i) *For $n \geq 1$ and $P \in \mathcal{P}_n$,*

$$\|P'w\|_{L_\infty[-1,1]} \leq CnT(a_{2n})^{1/2}\|Pw\|_{L_\infty[-1,1]}. \tag{1.14}$$

(ii) *For $n \geq 1$, $P \in \mathcal{P}_n$, and $|x| < a_n$,*

$$|P'w|(x) \leq \frac{Cn}{\sqrt{1-|x|/a_n}}\|Pw\|_{L_\infty[-1,1]}. \tag{1.15}$$

We remark that under mild additional conditions, which are satisfied for $W_{k,\alpha}$ for all $k \geq 0$, $\alpha > 0$,

$$nT(a_{2n})^{1/2} \sim Q'(a_{2n}), \qquad n \to \infty, \tag{1.16}$$

(see Lemma 3.2(ii) below), so one may reformulate (1.14) as

$$\|P'w\|_{L_\infty[-1,1]} \leq CQ'(a_{2n})\|Pw\|_{L_\infty[-1,1]}. \tag{1.17}$$

One may combine (1.14) and (1.15) as follows:

COROLLARY 1.3. *Let $w \in \mathcal{W}$. For $n \geq 1$, $P \in \mathcal{P}_n$, and $x \in [-1,1]$,*

$$|P'w|(x) \leq \frac{Cn}{|1-|x|/a_n|^{1/2} + T(a_{2n})^{-1/2}}\|Pw\|_{L_\infty[-1,1]}. \tag{1.18}$$

The above is the analogue of the classical consequence of (1.1) and (1.2):

$$|P'(x)| \leq \frac{Cn}{|1-|x||^{1/2}+n^{-1}}\|P\|_{L_\infty[-1,1]}, \quad P \in \mathcal{P}_n, \quad x \in (-1,1).$$

As examples of Theorem 1.2, we present the following.

COROLLARY 1.4. (i) *Let $\alpha > 0$ and $W_{0,\alpha}$ be given by (1.4). Then for $n \geq 1$ and $P \in \mathcal{P}_n$,*

$$\|P'W_{0,\alpha}\|_{L_\infty[-1,1]} \leq Cn^{(2\alpha+2)/(2\alpha+1)}\|PW_{0,\alpha}\|_{L_\infty[-1,1]}. \tag{1.19}$$

(ii) *Let $k \geq 1$, $\alpha > 0$, and $W_{k,\alpha}$ be given by (1.5). Then for $n \geq 1$ and $P \in \mathcal{P}_n$,*

$$\|P'W_{k,\alpha}\|_{L_\infty[-1,1]} \leq Cn\left[\prod_{j=1}^{k}\log_j n\right]^{1/2}(\log_k n)^{(\alpha+1)/2\alpha}\|PW_{k,\alpha}\|_{L_\infty[-1,1]}. \tag{1.20}$$

(iii) *Let $\alpha > 0$ and $w(x) := (1-x^2)^\alpha$. Then for $n \geq 1$ and $P \in \mathcal{P}_n$,*

$$\|P'w\|_{L_\infty[-1,1]} \leq Cn^2\|Pw\|_{L_\infty[-1,1]}. \tag{1.21}$$

We remark that under mild additional conditions involving Q''', which hold for $W_{k,\alpha}$, $k \geq 0$, $\alpha > 0$, we can show that (1.14) is sharp with respect to the dependence on n. The proof is lengthy, and involves analysis of L_∞ extremal polynomials for w. For the proof in a closely related situation, we refer the reader to [7, pp. 71–78]. Note that (1.21) is a classical inequality for ultraspherical weights [1].

Next, we turn to inequalities for $(Pw)'$. These are different from those for $P'w$, since for x close to a_n, $|(Pw)'|(x)$ admits a far better estimate than $|P'w|(x)$. A similar situation occurs for Erdös weights on $\mathbf{R}$ (cf. [6]).

THEOREM 1.5. *Let $w \in \mathcal{W}$. Then*
(a) *For $n \geq 1$, $P \in \mathcal{P}_n$ and $|x| < a_n$,*

$$(1.22) \qquad |(Pw)'(x)| \leq \frac{Cn}{(1-|x|/a_n)^{1/2} + T(a_{2n})^{-1/2}} \|Pw\|_{L_\infty[-1,1]};$$

(b) *For $n \geq 1$, $P \in \mathcal{P}_n$ and $|x| \leq a_n$,*

$$(1.23) \qquad |(Pw)'(x)| \leq CnT(a_{2n}) \left\{ \left(1 - \frac{|x|}{a_n}\right)^{1/2} + (nT(a_{2n}))^{-1/3} \right\} \|Pw\|_{L_\infty[-1,1]}.$$

In particular,

$$(1.24) \qquad |(Pw)'(a_n)| \leq C(nT(a_{2n}))^{2/3} \|Pw\|_{L_\infty[-1,1]}.$$

Note that for suitable A, and $|x|/a_n \geq 1 - AT(a_{2n})^{-1}$, (1.23) provides a better estimate than (1.22), while for $|x|/a_n < 1 - AT(a_{2n})^{-1}$, (1.22) provides a better estimate.

Next, we turn to estimates of *Christoffel functions.* Recall that the nth Christoffel function for w^2 is

$$(1.25) \qquad \lambda_n(w^2, x) := \inf_{P \in \mathcal{P}_{n-1}} \frac{1}{P^2(x)} \int_{-1}^{1} (Pw)^2(t)\,dt, \qquad x \in (-1,1).$$

It turns out that a particularly simple way to find upper bounds for λ_n^{-1} involves the Markov–Bernstein inequality for w. This idea has been used elsewhere [7].

THEOREM 1.6. *Let $w \in \mathcal{W}$. Then for $n \geq 1$,*

$$(1.26) \qquad \sup_{x \in [-1,1]} \lambda_n^{-1}(w^2, x) w^2(x) \leq CnT(a_{2n})^{1/2}.$$

We note that this result is sharp under mild additional conditions on Q; see [9] for the proof in a similar situation. As a corollary of the above estimate, we obtain Nikolskii inequalities.

COROLLARY 1.7. *Let $w \in \mathcal{W}$ and $0 < p < q \leq \infty$. Then for $n \geq 1$ and $P \in \mathcal{P}_n$,*

$$(1.27) \qquad \|Pw\|_{L_q[-1,1]} \leq C(nT(a_{2n})^{1/2})^{1/p - 1/q} \|Pw\|_{L_p[-1,1]}.$$

This paper is organized as follows. In §2, we present the basic ingredients of the proof—contour integral estimates, and integral equations with logarithmic kernel. In §3, we prove some technical lemmas. In §4, we present estimates for the measure $\mu_n(x)$, and in §5, we estimate the majorisation function $U_n(x+iy)$. In §6, we prove the Markov–Bernstein inequalities, and in §7, we prove Theorems 1.6 and Corollary 1.7.

2. The two basic ingredients. The first ingredient is potential theory and an integral equation used in majorisation of weighted polynomials.

LEMMA 2.1. *Let $w := e^{-Q} \in \mathcal{W}$. For $n \geq 1$, let $a_n = a_n(Q)$ be the root of* (1.8). *For $x \in (-1,1) \backslash \{0\}$, let*

$$(2.1) \qquad \mu_n(x) := \frac{2}{\pi^2} \int_0^1 \frac{\sqrt{1-x^2}}{\sqrt{1-s^2}} \frac{a_n s Q'(a_n s) - a_n x Q'(a_n x)}{n(s^2 - x^2)} ds.$$

Furthermore, let

$$\chi_n := \frac{2}{\pi}\int_0^1 \frac{Q(a_n t)}{\sqrt{1-t^2}}\,dt + n\log 2, \tag{2.2}$$

and for $z \in \mathbf{C}$ such that $|z| < 1/a_n$, let

$$U_n(z) := \int_{-1}^1 (\log|z-t|)\mu_n(t)\,dt - \frac{Q(a_n|z|)}{n} + \frac{\chi_n}{n}. \tag{2.3}$$

(a) *Then for almost every $x \in (-1,1)$,*

$$0 < \mu_n(x) < \infty, \tag{2.4}$$

$$\int_{-1}^1 \mu_n(x)dx = 1, \tag{2.5}$$

and

$$\int_{-1}^1 \frac{\mu_n(x)}{1-x}dx = \frac{Q'(a_n)}{n}. \tag{2.6}$$

(b) *Furthermore,*

$$U_n(x) = 0, \qquad x \in [-1,1], \tag{2.7}$$

and

$$(xU_n'(x))' < 0, \quad U_n(x) < 0, \quad U_n'(x) < 0, \quad x \in \left(1, \frac{1}{a_n}\right). \tag{2.8}$$

(c) *For $n \geq 1$, $P \in \mathcal{P}_n$ and $z \in \mathbf{C}$ such that $|z| < 1$,*

$$|P(z)w(|z|)| \leq \|Pw\|_{L_\infty[-a_n,a_n]} \exp\left(nU_n\left(\frac{z}{a_n}\right)\right). \tag{2.9}$$

Furthermore,

$$\|Pw\|_{L_\infty[-1,1]} = \|Pw\|_{L_\infty[-a_n,a_n]}, \tag{2.10}$$

and if P is not identically zero,

$$|Pw|(x) < \|Pw\|_{L_\infty[-a_n,a_n]}, \qquad |x| > a_n. \tag{2.11}$$

Proof. See [11, Lemmas 5.1, 5.2, pp. 28–34]. One sets $f(x) := Q(a_n x)$, $x \in [-1/a_n, 1/a_n]$ in [11].

(a) The constant B in [11, eqn. (5.4), p. 28] is zero in view of (1.8), and the function μ_n above is the function $g(f;t)$ or $L[f'](t)$ of [11]. Then (2.5) and (2.6) are, respectively, (5.6) and (5.23) in [11].

(b) The identity (2.7) follows from [11, eqn. (5.1), p. 28]. Furthermore, (5.25) and (5.26) in [11, p. 23] ensure that $U_n'(a_n) = U_n(a_n) = 0$ and the first relation in (2.8), which is (5.27) in [11, p. 32], then implies the other two.

(c) See [13, pp. 74–75], or see [11, p. 51]. □

Next, we turn to estimates for derivatives of weighted polynomials, derived via Cauchy's integral formula. This method has been used elsewhere [4], [6].

LEMMA 2.2. *Let $w \in \mathcal{W}$. Let $x \in (-1,1)$, $\epsilon \in (0, 1-|x|)$, and $P \in \mathcal{P}_n$ for some $n \geq 1$. Then*

$$(2.12) \qquad |(Pw)'(x)| \leq \epsilon^{-1} e^{\tau} \|Pw\|_{L_\infty[-1,1]} \max_{|t-x|=\epsilon} \exp\left(nU_n\left(\frac{t}{a_n}\right)\right),$$

where

$$(2.13) \qquad \tau := \begin{cases} Q'(3\epsilon)2\epsilon, & \text{if } |x| \leq 2\epsilon, \\ \left[\frac{Q'(|x|+\epsilon)}{|x|} + Q''(|x|+\epsilon)\right]2\epsilon^2 & \text{if } |x| > 2\epsilon. \end{cases}$$

Proof. Fix $x \in (-1,1)$ and define the entire function

$$\hat{w}(t) := \exp(-Q(x) - Q'(x)(t-x)), \qquad t \in \mathbf{C},$$

so that $\hat{w}^{(j)}(x) = w^{(j)}(x)$, $j = 0, 1$. Then

$$(2.14) \qquad \begin{aligned} |(Pw)'(x)| = |(P\hat{w})'(x)| &= \left|\frac{1}{2\pi i}\int_{|t-x|=\epsilon} \frac{(P\hat{w})(t)}{(t-x)^2}dt\right| \\ &\leq \frac{1}{\epsilon}\max_{|t-x|=\epsilon} |P\hat{w}|(t) \\ &\leq \frac{1}{\epsilon}\max_{|t-x|=\epsilon} |\hat{w}(t)/w(|t|)| \cdot \max_{|t-x|=\epsilon} |P(t)w(|t|)| \\ &\leq \frac{1}{\epsilon}\max_{|t-x|=\epsilon} |\hat{w}(t)/w(|t|)| \cdot \|Pw\|_{L_\infty[-1,1]} \cdot \max_{|t-x|=\epsilon} \exp(nU_n(t/a_n)), \end{aligned}$$

by (2.9). It remains to estimate $|\hat{w}(t)/w(|t|)|$. Suppose first that $|x| \leq 2\epsilon$. Then for $|t-x| = \epsilon$,

$$(2.15) \qquad \begin{aligned} |\hat{w}(t)/w(|t|)| &= \exp(-Q(x) - Q'(x)(\operatorname{Re} t - x) + Q(|t|)) \\ &\leq \exp(-Q(|x|) + Q'(2\epsilon)\epsilon + Q(|x|+\epsilon)), \end{aligned}$$

where we have used the monotonicity of Q and Q'. Finally,

$$Q(|x|+\epsilon) - Q(|x|) \leq Q'(|x|+\epsilon)\epsilon \leq Q'(3\epsilon)\epsilon,$$

and (2.13) follows for $|x| \leq 2\epsilon$.

Next, suppose $x > 2\epsilon$. Then

$$(2.16) \qquad \begin{aligned} -Q(x) - Q'(x)(\operatorname{Re} t - x) + Q(|t|) &= Q(|t|) - Q(\operatorname{Re} t) \\ &\quad + Q(\operatorname{Re} t) - Q(x) - Q'(x)(\operatorname{Re} t - x) \\ &= Q'(\xi)(|t| - \operatorname{Re} t) + \tfrac{1}{2}Q''(\eta)(\operatorname{Re} t - x)^2, \end{aligned}$$

where ξ lies between $\operatorname{Re} t$ and $|t|$, and η lies between $\operatorname{Re} t$ and x. Here $\xi, \eta \in [x-\epsilon, x+\epsilon]$. Furthermore, by the elementary inequality $(a^2+b^2)^{1/2} \leq a + b^2/a$, for $a, b > 0$,

$$|t| - \operatorname{Re} t \leq \frac{(\operatorname{Im} t)^2}{\operatorname{Re} t} \leq \frac{\epsilon^2}{x-\epsilon} \leq \frac{2\epsilon^2}{x},$$

so

$$Q'(\xi)(|t| - \operatorname{Re} t) \le \frac{2Q'(x+\epsilon)\epsilon^2}{x}. \tag{2.17}$$

Furthermore, $uQ''(u) = Q'(u)(T(u) - 1)$ is increasing in (0,1), so

$$\eta Q''(\eta) \le (x+\epsilon)Q''(x+\epsilon),$$

and hence

$$Q''(\eta) \le \frac{x+\epsilon}{x-\epsilon}Q''(x+\epsilon) \le 3Q''(x+\epsilon).$$

Then

$$\tfrac{1}{2}Q''(\eta)(\operatorname{Re} t - x)^2 \le \tfrac{3}{2}\epsilon^2 Q''(x+\epsilon). \tag{2.18}$$

Combining (2.14) to (2.18) yields (2.13) for $x \ge 2\epsilon$, and the case $x \le -2\epsilon$ is similar. □

3. Technical lemmas. In this section, we present some elementary consequences of the hypothesis $w \in \mathcal{W}$.

LEMMA 3.1. *Let* $w = e^{-Q} \in \mathcal{W}$.

(i) *We have for* $0 < x \le Lx < 1$,

$$L^{T(x)} \le \frac{LxQ'(Lx)}{xQ'(x)} \le L^{T(Lx)}. \tag{3.1}$$

(ii) $Q'(x)$ *and* $xQ''(x)$ *are increasing in* $(0,1)$.

(iii)

$$\frac{xQ''(x)}{Q'(x)} \sim T(x) \quad \textit{in } (0,1]. \tag{3.2}$$

(iv)

$$\frac{Q'(x)}{x} \le (T(0+) - 1)^{-1}Q''(x) \quad \textit{in } (0,1], \tag{3.3}$$

and

$$Q'(0) = 0. \tag{3.4}$$

(v)

$$\int_0^{1/2} \frac{Q'(x)}{x}dx < \infty. \tag{3.5}$$

Proof. (i) Now

$$\frac{LxQ'(Lx)}{xQ'(x)} = \exp\left(\int_x^{Lx} \frac{d}{dt}\log[tQ'(t)]dt\right) = \exp\left(\int_x^{Lx} \frac{T(t)}{t}dt\right).$$

Here the monotonicity of T ensures that

$$T(x)\log L \le \int_x^{Lx} \frac{T(t)}{t}dt \le T(Lx)\log L.$$

Then (3.1) follows.

(ii) Since

$$xQ''(x) = (T(x) - 1)Q'(x), \tag{3.6}$$

the monotonicity of T and Q' yields the monotonicity of $xQ''(x)$.

(iii) By (3.6), for $x \in (0,1)$,

$$xQ''(x) = T(x)Q'(x)\left(1 - \frac{1}{T(x)}\right)\begin{cases} \le T(x)Q'(x) \\ \ge T(x)Q'(x)\left(1 - \frac{1}{T(0+)}\right). \end{cases}$$

(iv) Firstly, (3.6), the monotonicity of T and the fact that $T(0+) > 1$ yield (3.3). Next, the evenness of Q and continuity of Q' force (3.4).

(v) If $0 < \delta < \frac{1}{2}$, inequality (3.3) yields

$$\int_\delta^{1/2} \frac{Q'(s)}{s} ds \le (T(0+) - 1)^{-1}\left[Q'(1/2) - Q'(\delta)\right].$$

Now let $\delta \to 0+$. □

LEMMA 3.2. *Let $w = e^{-Q} \in \mathcal{W}$ and $a_n = a_n(Q)$, $n \ge 1$.*

(i) *For $j = 1, 2$ and n large enough,*

$$Q^{(j)}(a_n) = O(nT(a_n)^{j-1/2}). \tag{3.7}$$

(ii) *If also*

$$\frac{Q'(x)}{Q(x)} \sim T(x), \qquad x \text{ near } 1, \tag{3.8}$$

then for $j = 0, 1, 2$, and n large enough,

$$Q^{(j)}(a_n) \sim nT(a_n)^{j-1/2}. \tag{3.9}$$

(iii) *We have*

$$Q'(a_n) = O(n^2); \tag{3.10}$$

$$T(a_n) = O(n^2). \tag{3.11}$$

(iv) *For $t \in (0, \infty)$,*

$$\frac{1}{tT(0+)} \ge \frac{a_t'}{a_t} \ge \frac{1}{tT(a_t)}. \tag{3.12}$$

(v) *For $u \in (0, \infty)$ and $r \ge 1$,*

$$\frac{a_{ru}}{a_u} \ge 1 + \frac{\log r}{T(a_{ru})}. \tag{3.13}$$

Proof.

$$\text{(i)} \qquad \frac{n}{a_nQ'(a_n)} = \frac{2}{\pi}\int_0^1 \frac{a_ntQ'(a_nt)}{a_nQ'(a_n)}\frac{1}{\sqrt{1-t^2}}dt$$

$$\begin{aligned}&\geq \frac{2}{\pi}\int_0^1 t^{T(a_n)}\frac{1}{\sqrt{1-t^2}}dt \quad \text{(by (3.1))}\\ (3.14)\qquad &\geq \frac{2}{\pi}\left(1-\frac{1}{T(a_n)}\right)^{T(a_n)}\int_{1-1/T(a_n)}^1 \frac{1}{\sqrt{1-t^2}}dt\\ &\geq C\left(1-\frac{1}{T(0+)}\right)^{T(0+)} T(a_n)^{-1/2}.\end{aligned}$$

Then (3.7) follows for $j=1$, and (3.2) implies it for $j=2$.

(ii) From (3.14),

$$\begin{aligned}\frac{n}{a_nQ'(a_n)} &\leq \frac{2}{\pi}\int_0^{1-1/T(a_n)}\frac{Q'(a_nt)}{Q'(a_n)}\frac{1}{\sqrt{1-t^2}}dt+\frac{2}{\pi}\int_{1-1/T(a_n)}^1\frac{1}{\sqrt{1-t^2}}dt\\ &\leq C_1T(a_n)^{1/2}\int_0^1\frac{Q'(a_nt)}{Q'(a_n)}dt+C_1T(a_n)^{-1/2}\\ &= C_1T(a_n)^{1/2}\frac{Q(a_n)-Q(0)}{a_nQ'(a_n)}+C_1T(a_n)^{-1/2}\\ &\leq C_2T(a_n)^{1/2}\frac{Q(a_n)}{a_nQ'(a_n)}+C_1T(a_n)^{-1/2}\leq C_3T(a_n)^{-1/2},\end{aligned}$$

n large enough, by (3.8) and monotonicity of Q. Hence

$$a_nQ'(a_n)\sim nT(a_n)^{1/2}.$$

Then (3.2) yields (3.9) for $j=2$, and (3.8) yields it for $j=0$.

(iii) From (3.7) and (1.13),

$$Q'(a_n)=O(nT(a_n)^{1/2})=O(nQ'(a_n)^{1/2}),$$

whence (3.10) follows. Then (1.13) yields (3.11).

(iv) Differentiating (1.8) with respect to u yields

$$\begin{aligned}1&=\frac{2}{\pi}\int_0^1\frac{d}{du}(a_utQ'(a_ut))\frac{dt}{\sqrt{1-t^2}}\\ &=\frac{2}{\pi}\int_0^1 T(a_ut)Q'(a_ut)a_u't\frac{dt}{\sqrt{1-t^2}}\\ &=\frac{a_u'}{a_u}\frac{2}{\pi}\int_0^1 T(a_ut)a_utQ'(a_ut)\frac{dt}{\sqrt{1-t^2}}.\end{aligned}$$

Since

$$T(0+)\leq T(a_ut)\leq T(a_u),\qquad t\in(0,1],$$

we obtain (3.12) from (1.8).

(v)

$$\frac{a_{ru}}{a_u} = \exp\left(\int_u^{ru} \frac{a_t'}{a_t} dt\right) \;\geq\; \exp\left(\int_u^{ru} \frac{dt}{tT(a_t)}\right)$$

$$\geq \exp\left(\frac{\log r}{T(a_{ru})}\right) \;\geq\; 1 + \frac{\log r}{T(a_{ru})}. \qquad \square$$

LEMMA 3.3. *Let $w \in \mathcal{W}$.*

(i) *For $0 < \alpha < \Delta$, we have*

$$\frac{a_{\Delta n} Q'(a_{\Delta n})}{a_{\alpha n} Q'(a_{\alpha n})} \geq \frac{\Delta}{\alpha} > 1, \qquad n \geq 1. \tag{3.15}$$

(ii) *For $n \geq 1$ and $x \in [-1, 1]$,*

$$a_n x Q'(a_n x)(1 - |x|)^{1/2} \leq Cn. \tag{3.16}$$

Furthermore, for $n \geq 1$ and $s \in [-a_n, a_n]$,

$$sQ'(s)\left(1 - \frac{|s|}{a_n}\right)^{1/2} \leq Cn. \tag{3.17}$$

Proof.

(i)

$$\frac{a_{\Delta n} Q'(a_{\Delta n})}{a_{\alpha n} Q'(a_{\alpha n})} = \exp\left(\int_{\alpha n}^{\Delta n} \frac{d}{dt}[\log(a_t Q'(a_t))] dt\right)$$

$$= \exp\left(\int_{\alpha n}^{\Delta n} T(a_t)\frac{a_t'}{a_t} dt\right) \geq \exp\left(\int_{\alpha n}^{\Delta n} \frac{dt}{t}\right) = \frac{\Delta}{\alpha},$$

by (3.12).

(ii) Now uniformly for $|x| \in [0, 1)$,

$$\int_{|x|}^1 \frac{ds}{\sqrt{1 - s^2}} \sim (1 - |x|)^{1/2}. \tag{3.18}$$

Hence for $|x| \in [0, 1)$,

$$a_n x Q'(a_n x)(1 - |x|)^{1/2} = a_n |x| Q'(a_n |x|)(1 - |x|)^{1/2}$$

$$\leq C a_n |x| Q'(a_n |x|) \frac{2}{\pi} \int_{|x|}^1 \frac{ds}{\sqrt{1 - s^2}}$$

$$\leq C \frac{2}{\pi} \int_{|x|}^1 a_n s Q'(a_n s) \frac{ds}{\sqrt{1 - s^2}} \leq Cn.$$

For $|x| = 1$, (3.16) is trivial. Setting $x = s/a_n$ yields (3.17). $\square$

4. Estimates for the measure μ_n. In this section, we present some estimates for the measure $\mu_n(x)$ that are of independent interest. The methods we use are similar to those in [4], [6], [9].

THEOREM 4.1. *Let $w \in \mathcal{W}$. Then for $n \geq 1$,*

$$\text{(a)} \quad \max_{x\in[-1,1]} \mu_n(x)\sqrt{1-x^2} \leq C; \tag{4.1}$$

$$\text{(b)} \quad \max_{x\in[-1,1]} \mu_n(x)/\sqrt{1-x^2} \leq CT(a_n). \tag{4.2}$$

As a corollary, we shall deduce the following.

COROLLARY 4.2. *For $n \geq 1$,*

$$\max_{x\in[-1,1]} \mu_n(x) \leq CT(a_n)^{1/2}. \tag{4.3}$$

Under mild additional conditions, we can show the following.

THEOREM 4.3. *In addition to $w \in \mathcal{W}$, let us assume* (3.8). *Fix $\beta \in (0,1)$. Then we have*

$$\sqrt{1-x^2}\mu_n(x) \geq C, \quad |x| \leq \frac{a_{\beta n}}{a_n}, \quad n \geq 1. \tag{4.4}$$

We note that under mild additional conditions, we can show that (4.2) and (4.3) are sharp. See, respectively, [5] and [8] for the sharpness of (4.2) and (4.3) in the related situation of Erdös weights. We turn now to the proofs of the above results.

Proof of Theorem 4.1(a). Since μ_n is even, we may consider $x \in (0,1)$. For $n \geq 1$, and s, $x \in (0,1)$, set

$$\Delta_n(s,x) := \frac{a_n sQ'(a_n s) - a_n xQ'(a_n x)}{a_n s - a_n x}. \tag{4.5}$$

Then

$$\sqrt{1-x^2}\mu_n(x) = \frac{2a_n}{n\pi^2}(1-x^2)\int_0^1 \frac{\Delta_n(s,x)}{s+x}\frac{ds}{\sqrt{1-s^2}}. \tag{4.6}$$

Note that since $(uQ'(u))' = Q'(u)T(u)$ is increasing in (0,1), $uQ'(u)$ is convex in (0,1). Hence for each fixed x, $\Delta_n(s,x)$ is an increasing function of $s \in (0,1)$.

Case I. $x \in [0, \frac{1}{4}]$. Now suppose that $x \in [0, \frac{1}{4}]$. Then $s \in (2x, 1]$ implies

$$\Delta_n(s,x) \leq \frac{a_n sQ'(a_n s)}{a_n s - a_n s/2} = 2Q'(a_n s).$$

Furthermore, $s \in (0, 2x]$ implies

$$\Delta_n(s,x) \leq \Delta_n(2x,x) \leq \frac{a_n 2xQ'(a_n 2x)}{a_n 2x - a_n x} = 2Q'(a_n 2x).$$

Hence for $0 \leq x \leq \frac{1}{4}$,

$$\begin{aligned}\sqrt{1-x^2}\mu_n(x) &\leq \frac{C_1}{n}(1-|x|)\left[\int_0^{2x}\frac{Q'(a_n 2x)}{s+x}ds + \int_{2x}^1 \frac{2Q'(a_n s)}{s}\frac{ds}{\sqrt{1-s^2}}\right] \\ &\leq \frac{C_2}{n}(1-|x|)\left[2Q'(1/2) + \int_{2x}^{1/2}\frac{Q'(a_n s)}{s}ds + \int_{1/2}^1 \frac{a_n sQ'(a_n s)}{\sqrt{1-s^2}}ds\right] \\ &\leq C,\end{aligned}$$

by Lemma 3.1(v) and the definition of a_n.

Case II. $x \in [\frac{1}{4}, 1)$. Next, suppose $\frac{1}{4} \leq x < 1$. Let

$$\delta(x) := \frac{1-x}{4}.$$

From (4.6) and as $x + 4\delta(x) = 1$, we have

$$\begin{aligned}\sqrt{1-x^2}\mu_n(x) &\leq \frac{C}{n}(1-x)\int_0^1 \frac{\Delta_n(s,x)}{\sqrt{1-s^2}}ds \\ &= \frac{C}{n}4\delta(x)\left[\int_0^{x-\delta(x)} + \int_{x-\delta(x)}^{x+\delta(x)} + \int_{x+\delta(x)}^1\right]\frac{\Delta_n(s,x)}{\sqrt{1-s^2}}ds \qquad (4.7) \\ &=: \frac{4C}{n}(I_1 + I_2 + I_3).\end{aligned}$$

Firstly,

$$\begin{aligned}I_1 &\leq \delta(x)\int_0^{x-\delta(x)} \frac{a_n x Q'(a_n x)}{a_n x - a_n s}\frac{ds}{\sqrt{1-s^2}} \\ &\leq C_1\delta(x)xQ'(a_n x)\int_0^{x-\delta(x)}\frac{ds}{(x-s)^{3/2}} \quad (\text{as } 1-s \geq x-s > 0) \qquad (4.8) \\ &\leq C_2 Q'(a_n x)\delta(x)^{1/2} \leq C_3 n,\end{aligned}$$

by (3.16). Next, if $s \in [x-\delta(x), x+\delta(x)]$, there exists ξ between s and x such that

$$\begin{aligned}\Delta_n(s,x) &= \frac{d}{du}(uQ'(u))\Big|_{u=a_n\xi} = T(a_n\xi)Q'(a_n\xi) \\ &\leq T(a_n[x+\delta(x)])Q'(a_n[x+\delta(x)]) \\ &\leq \delta(x)^{-1}\int_{x+\delta(x)}^{x+2\delta(x)} T(a_n s)Q'(a_n s)ds \\ &= (a_n\delta(x))^{-1}\int_{x+\delta(x)}^{x+2\delta(x)} \frac{d}{ds}(a_n sQ'(a_n s))ds \\ &\leq (a_n\delta(x))^{-1}a_n[x+2\delta(x)]Q'(a_n[x+2\delta(x)]) \\ &\leq C_4\delta(x)^{-3/2}a_n[x+2\delta(x)]Q'(a_n[x+2\delta(x)])\int_{x+2\delta(x)}^1\frac{ds}{\sqrt{1-s^2}} \\ &\qquad (\text{observe } 1-[x+2\delta(x)] = 2\delta(x)) \\ &\leq C_4\delta(x)^{-3/2}\int_{x+2\delta(x)}^1 \frac{a_n sQ'(a_n s)}{\sqrt{1-s^2}}ds \\ &\leq C_5\delta(x)^{-3/2}n.\end{aligned}$$

Then

$$\begin{aligned} I_2 &\leq \delta(x)\int_{x-\delta(x)}^{x+\delta(x)} C_5\delta(x)^{-3/2}n\frac{ds}{\sqrt{1-s^2}} \\ &\leq C_5\delta(x)^{-1/2}n\int_{x-\delta(x)}^{1}\frac{ds}{\sqrt{1-s^2}} \leq C_6 n, \end{aligned} \tag{4.9}$$

as $1-(x-\delta(x)) = 5\delta(x)$. Finally,

$$I_3 \leq \delta(x)\int_{x+\delta(x)}^{1}\frac{a_n sQ'(a_n s)}{a_n\delta(x)}\frac{ds}{\sqrt{1-s^2}} \leq Cn, \tag{4.10}$$

by the definition of a_n. Combining (4.7) to (4.10) yields

$$\sqrt{1-x^2}\mu_n(x) \leq C, \qquad x \in [\tfrac{1}{4}, 1). \qquad \square$$

Proof of Theorem 4.1(b). With the notation (4.5), and by (4.6), we have for $1 > |x| \geq \frac{1}{2}$,

$$\begin{aligned} \mu_n(x)/\sqrt{1-x^2} &\leq \frac{C}{n}\int_0^1 \frac{\Delta_n(s,x)}{\sqrt{1-s^2}}ds \\ &\leq \frac{C}{n}\int_0^1 \frac{\Delta_n(s,1)}{\sqrt{1-s^2}}ds, \end{aligned} \tag{4.11}$$

since $\Delta_n(s,x)$ is increasing in x for fixed s (as in the proof of Theorem 4.1(a)). Now let

$$\begin{aligned} A_n := a_n\int_0^1\frac{\Delta_n(s,1)}{\sqrt{1-s^2}}ds &= \int_0^1 \frac{a_nQ'(a_n)-a_nsQ'(a_ns)}{(1-s)^{3/2}(1+s)^{1/2}}ds \\ &\leq \int_0^1 \frac{a_nQ'(a_n)-a_nsQ'(a_ns)}{(1-s)^{3/2}}ds. \end{aligned}$$

Integrating by parts, we have

$$\begin{aligned} A_n &\leq \frac{2}{(1-s)^{1/2}}(a_nQ'(a_n)-a_nsQ'(a_ns))\Bigg|_{s=0}^{s=1} \\ &\quad -\int_0^1\frac{2}{(1-s)^{1/2}}\frac{d}{ds}(a_nQ'(a_n)-a_nsQ'(a_ns))ds \\ &= -2a_nQ'(a_n)+2a_n\int_0^1\frac{T(a_ns)Q'(a_ns)}{(1-s)^{1/2}}ds \\ &\leq 4T(a_n)\int_{1/2}^1\frac{Q'(a_ns)}{\sqrt{1-s}}ds \quad \text{(by the monotonicity of } Q' \text{ and } (1-s)^{-1/2}\text{)} \\ &\leq CT(a_n)a_n^{-1}\int_{1/2}^1\frac{a_nsQ'(a_ns)}{\sqrt{1-s^2}}ds \leq C_1T(a_n)n. \end{aligned}$$

Then from (4.11), we obtain for $\frac{1}{2} \leq |x| \leq 1$,

$$\mu_n(x)/\sqrt{1-x^2} \leq C_2 \frac{a_n}{n} A_n \leq C_3 T(a_n).$$

Since (4.1) implies that for $|x| \leq \frac{1}{2}$,

$$\mu_n(x) \leq C_4 \leq C_5 T(a_n),$$

we have our result. □

Proof of Corollary 4.2. For $x^2 \leq 1 - 1/T(a_n)$, (4.1) implies

$$\mu_n(x) \leq \frac{C}{\sqrt{1-x^2}} \leq C_1 T(a_n)^{1/2}.$$

For $1 \geq x^2 \geq 1 - 1/T(a_n)$, inequality (4.2) implies

$$\mu_n(x) \leq CT(a_n)\sqrt{1-x^2} \leq CT(a_n)^{1/2}. \quad \square$$

Finally, we turn to the following.

Proof of Theorem 4.3. Let $0 < \beta < \Delta < 1$. Now from (2.1), and the positivity of

$$\frac{a_n s Q'(a_n s) - a_n x Q'(a_n x)}{a_n s - a_n x},$$

we see that for $|x| \leq a_{\beta n}/a_n$,

$$\sqrt{1-x^2}\mu_n(x) \geq \frac{2}{n\pi^2} \int_{a_{\Delta n}/a_n}^{1} \frac{1-x^2}{s^2-x^2} \frac{a_n s Q'(a_n s) - a_n x Q'(a_n x)}{\sqrt{1-s^2}} ds.$$

Now for $s \in [a_{\Delta n}/a_n, 1]$, and $|x| \leq a_{\beta n}/a_n$,

$$\begin{aligned}\frac{a_n s Q'(a_n s) - a_n x Q'(a_n x)}{a_n Q'(a_n)} &= \frac{a_n s Q'(a_n s)}{a_n Q'(a_n)} \left[1 - \frac{a_n x Q'(a_n x)}{a_n s Q'(a_n s)}\right] \\ &\geq s^{T(a_n)} \left[1 - \frac{a_{\beta n} Q'(a_{\beta n})}{a_{\Delta n} Q'(a_{\Delta n})}\right] \quad \text{(by (3.1))} \\ &\geq s^{T(a_n)} \left[1 - \frac{\beta}{\Delta}\right],\end{aligned}$$

by (3.15). Furthermore, as $s^2 > x^2$ for $s \in [a_{\Delta n}/a_n, 1]$,

$$1 - x^2 \geq s^2 - x^2 > 0.$$

Hence

$$\sqrt{1-x^2}\mu_n(x) \geq \frac{2}{n\pi^2} \left[1 - \frac{\beta}{\Delta}\right] a_n Q'(a_n) \int_{a_{\Delta n}/a_n}^{1} s^{T(a_n)} \frac{ds}{\sqrt{1-s^2}}.$$

Now by (3.13), for $n \geq 1$,

$$\frac{a_{\Delta n}}{a_n} \leq 1 - \frac{C}{T(a_n)}.$$

Hence if δ is small enough,

$$\sqrt{1-x^2}\mu_n(x) \geq \frac{2}{n\pi^2}\left[1-\frac{\beta}{\Delta}\right] a_n Q'(a_n) \int_{1-\delta/T(a_n)}^{1} s^{T(a_n)} \frac{ds}{\sqrt{1-s^2}}$$
$$\geq CT(a_n)^{1/2}\left(1-\frac{\delta}{T(a_n)}\right)^{T(a_n)} \int_{1-\delta/T(a_n)}^{1} \frac{ds}{\sqrt{1-s^2}}$$
$$\geq C_1,$$

by (3.9). □

We remark that we used (3.8) only in the last two lines of the above proof, in estimating $a_nQ'(a_n)$ from below.

5. Estimates for the majorisation function U_n. In this section, we estimate the quantity

$$\max_{|t-x|=\epsilon} \exp\left(nU_n\left(\frac{t}{a_n}\right)\right),$$

which appears in Lemma 2.2. Recall that the majorisation function is defined by (2.3).

THEOREM 5.1. *Let $w := e^{-Q} \in \mathcal{W}$. Let $0 < \beta < 1$, and for a fixed $\eta > 0$, let*

$$\epsilon_n := \eta(nT(a_n)^{1/2})^{-1}, \qquad n \geq 1. \tag{5.1}$$

Then

$$\max_{|t-x|\leq\epsilon_n} \exp\left(nU_n\left(\frac{t}{a_n}\right)\right) \leq C, \tag{5.2}$$

uniformly for $n \geq 1$ and for real x satisfying

$$|x| \leq a_{\beta n}. \tag{5.3}$$

Proof. Let $|t-x| \leq \epsilon_n$. Then we can write $t/a_n = a+ib$, where

$$|a| \leq \frac{|x|+\epsilon_n}{a_n} \leq \frac{a_{\beta n} + \eta/[nT(a_n)^{1/2}]}{a_n}$$
$$\leq \frac{a_{\beta n}(1+C\eta/T(a_n))}{a_n},$$

where in the last inequality we used the fact that $n^{-1} \leq C_1T(a_n)^{-1/2}$, which follows from (3.11). Now, if η is small enough, we get from (3.13) that $a_{\beta n}(1+C\eta/T(a_n))/a_n < 1$, and so

$$|x| + \epsilon_n < a_n. \tag{5.4}$$

Also,

$$|b| \leq \frac{\epsilon_n}{a_n} \leq \frac{\epsilon_n}{a_1}. \tag{5.5}$$

Then since $|a| < 1$, $U_n(a) = 0$ (see (2.7)); so we have from (2.3),

$$\begin{aligned}
U_n(t/a_n) &= U_n(a+ib) - U_n(a) \\
&= \int_{-1}^{1} \log\left|\frac{a+ib-t}{a-t}\right| \mu_n(t)dt - \frac{Q(a_n[a^2+b^2]^{1/2})}{n} + \frac{Q(a_n|a|)}{n} \\
&\leq \frac{1}{2}\int_{-1}^{1} \log\left[1+\left(\frac{b}{a-t}\right)^2\right] \mu_n(t)dt \\
&= \frac{1}{2}\int_{-1}^{1} \log\left[1+\left(\frac{|b|}{|a|-t}\right)^2\right] \mu_n(t)dt \quad \text{(as } \mu_n \text{ is even)} \\
&\leq \int_{0}^{1} \log\left[1+\left(\frac{|b|}{|a|-t}\right)^2\right] \mu_n(t)dt \qquad (5.6) \\
&\leq C_1 T(a_n)^{1/2} \int_{0}^{1} \log\left[1+\left(\frac{|b|}{|a|-t}\right)^2\right] dt \quad \text{(by Corollary 4.2)} \\
&= C_1 T(a_n)^{1/2}|b| \int_{(|a|-1)/|b|}^{|a|/|b|} \log\left[1+\frac{1}{s^2}\right] ds \quad \text{(substituting } |a|-t = s|b|) \\
&\leq C_1 T(a_n)^{1/2}\left(\frac{\epsilon_n}{a_1}\right) \int_{-\infty}^{\infty} \log\left[1+\frac{1}{s^2}\right] ds \quad \text{(by (5.5))} \\
&\leq \frac{C_2}{n},
\end{aligned}$$

by the choice (5.1) of ϵ_n. Hence $\exp(nU_n(t/a_n)) \leq \exp(C_2)$. □

Note that in the above result, we used the bound in Corollary 4.2 for μ_n. Next, we use the bound of Theorem 4.1(a).

THEOREM 5.2. *Let $w \in \mathcal{W}$. Let $n \geq 1$, and for real x satisfying*

$$|x| \leq a_n\left(1 - \frac{4}{n^2 a_1^2}\right), \tag{5.7}$$

let

$$\epsilon_n(x) := \frac{1}{n}\left(1 - \frac{|x|}{a_n}\right)^{1/2}. \tag{5.8}$$

Then

$$\max_{|t-x|\leq\epsilon_n(x)} \exp\left(nU_n\left(\frac{t}{a_n}\right)\right) \leq C. \tag{5.9}$$

Proof. Let x satisfy (5.7), and for $|t-x| \leq \epsilon_n(x)$, write $t/a_n = a + ib$. Then

$$\begin{aligned}|a| &\le \frac{|x|+\epsilon_n(x)}{a_n}\\ &= 1-\frac{1-|x|}{a_n}+\frac{\epsilon_n(x)}{a_n}\\ &\le 1+\epsilon_n(x)\left[-n^2\epsilon_n(x)+\frac{1}{a_1}\right] \quad \text{(by (5.8))}\\ &\le 1-\frac{n^2\epsilon_n^2(x)}{2},\end{aligned} \tag{5.10}$$

since (5.7) and (5.8) ensure that $\epsilon_n(x) \ge 2/(n^2 a_1)$. Furthermore,

$$|b| \le \frac{\epsilon_n(x)}{a_1}. \tag{5.11}$$

Since (2.7) shows that $U_n(a) = 0$, we obtain from (5.6),

$$\begin{aligned}U_n\left(\frac{t}{a_n}\right) &= U_n(a+ib)-U_n(a)\\ &\le \int_0^1 \log\left[1+\left(\frac{|b|}{|a|-t}\right)^2\right]\mu_n(t)dt\\ &\le C\int_0^1 \log\left[1+\left(\frac{|b|}{|a|-t}\right)^2\right]\frac{dt}{\sqrt{1-t}} \quad \text{(by Theorem 4.1(a))}\\ &= C|b|\int_{(|a|-1)/|b|}^{|a|/|b|}\log\left[1+\frac{1}{s^2}\right]\frac{ds}{\sqrt{1-|a|+s|b|}},\end{aligned}$$

by the substitution $|a|-t = s|b|$. Now let

$$\delta := \delta(a) := \frac{1-|a|}{2}. \tag{5.12}$$

We write

$$U_n\left(\frac{t}{a_n}\right) \le C|b|\left(\int_{-2\delta/|b|}^{-\delta/|b|}+\int_{-\delta/|b|}^{|a|/|b|}\right)\log\left[1+\frac{1}{s^2}\right]\frac{ds}{\sqrt{2\delta+s|b|}} =: I_1+I_2. \tag{5.13}$$

Then

$$\begin{aligned}I_1 &= C|b|^{1/2}\int_{-2\delta/|b|}^{-\delta/|b|}\log\left[1+\frac{1}{s^2}\right]\frac{ds}{\sqrt{2\delta/|b|+s}}\\ &\le C|b|^{1/2}\log\left[1+\left(\frac{|b|}{\delta}\right)^2\right]\int_0^{\delta/|b|}u^{-1/2}du\end{aligned} \tag{5.14}$$

$$= C|b|^{1/2}\log\left[1+\left(\frac{|b|}{\delta}\right)^2\right]2\left(\frac{\delta}{|b|}\right)^{1/2}$$
$$\leq 4C|b|\delta^{-1/2},$$

where we have used the inequality $\log[1+u^2] \leq 2u$, $u \in [0,\infty)$.

Next,

$$I_2 = C|b|\int_{-\delta/|b|}^{|a|/|b|}\log\left[1+\frac{1}{s^2}\right]\frac{ds}{\sqrt{2\delta+s|b|}}$$
$$\leq C|b|\delta^{-1/2}\int_{-\delta/|b|}^{|a|/|b|}\log\left[1+\frac{1}{s^2}\right]ds \tag{5.15}$$
$$\leq C|b|\delta^{-1/2}\int_{-\infty}^{\infty}\log\left[1+\frac{1}{s^2}\right]ds.$$

So from (5.13) to (5.15) and (5.11),

$$nU_n\left(\frac{t}{a_n}\right) \leq nC_1|b|\delta^{-1/2} \leq nC_1\epsilon_n(x)\delta^{-1/2} \leq C_2,$$

since (5.12) and (5.10) show that

$$\delta = \frac{1-|a|}{2} \geq \frac{n^2\epsilon_n^2(x)}{4}. \qquad \square$$

Finally, we use the bound of Theorem 4.1(b).

THEOREM 5.3. *Let $w \in \mathcal{W}$ and let $\eta > 0$ be fixed. For $n \geq 1$, and $|x| \leq a_n$, let*

$$\epsilon_n(x) := \eta\min\left\{(nT(a_n))^{-2/3}, \left(nT(a_n)\sqrt{1-|x|/a_n}\right)^{-1}\right\}. \tag{5.16}$$

Then if η is small enough (independently of x),

$$\max_{|t-x|\leq\epsilon_n(x)}\exp\left(nU_n\left(\frac{t}{a_n}\right)\right) \leq C. \tag{5.17}$$

Proof. Let $|x| \leq a_n$ and for $|t-x| \leq \epsilon_n(x)$, write $t/a_n = a+ib$. Here,

$$|b| \leq \frac{\epsilon_n(x)}{a_n} \leq \frac{\epsilon_n(x)}{a_1}. \tag{5.18}$$

Furthermore, by (3.11),

$$nT(a_n) \geq C_1T(a_n)^{3/2},$$

so

$$\epsilon_n(x) \leq \eta(nT(a_n))^{-2/3} \leq \eta C_2T(a_n)^{-1},$$

and hence

$$|x|+\epsilon_n(x) \leq a_n\left(1+\frac{C_3\eta}{T(a_n)}\right) \leq a_{2n}, \tag{5.19}$$

if η is small enough. In particular, then

$$\frac{|t|}{a_n} \le \frac{|x| + \epsilon_n(x)}{a_n} \le \frac{a_{2n}}{a_n} < 1. \tag{5.20}$$

Thus $U_n(t/a_n)$ is well defined for $|t - x| \le \epsilon_n(x)$. We now consider the cases $|a| \le 1$ and $|a| > 1$ separately.

Case I. $|a| \le 1$. Then $U_n(a) = 0$, and by (5.6),

$$\begin{aligned}
U_n\left(\frac{t}{a_n}\right) &= U_n(a+ib) - U_n(a)\\
&\le \int_0^1 \log\left[1 + \left(\frac{|b|}{|a|-t}\right)^2\right] \mu_n(t)\,dt\\
&\le CT(a_n)\int_0^1 \log\left[1 + \left(\frac{|b|}{|a|-t}\right)^2\right] \sqrt{1-t}\,dt \quad \text{(by Theorem 4.1(b))}\\
&= CT(a_n)|b| \int_{(|a|-1)/|b|}^{|a|/|b|} \log\left[1+\frac{1}{s^2}\right](1-|a|+s|b|)^{1/2}ds\\
&\qquad \text{(substituting } |a| - t = s|b|)\\
&\le CT(a_n)|b| \int_{-\infty}^{\infty} \log\left[1+\frac{1}{s^2}\right]\{(1-|a|)^{1/2} + |sb|^{1/2}\}\,ds\\
&\qquad \text{(by the inequality } (u+v)^{1/2} \le |u|^{1/2} + |v|^{1/2},\ u+v\ge 0)\\
&= CT(a_n)|b|(1-|a|)^{1/2}\int_{-\infty}^{\infty} \log\left[1+\frac{1}{s^2}\right] ds\\
&\qquad +CT(a_n)|b|^{3/2}\int_{-\infty}^{\infty} \log\left[1+\frac{1}{s^2}\right]|s|^{1/2}ds \qquad (5.21)\\
&\le C_1\left\{T(a_n)\epsilon_n(x)\left(1-\frac{|x|}{a_n}+\frac{\epsilon_n(x)}{a_n}\right)^{1/2} + T(a_n)\epsilon_n(x)^{3/2}\right\}\\
&\qquad \left(\text{by choice, } |a| \ge \tfrac{|x|}{a_n} - \tfrac{\epsilon_n(x)}{a_n}\right)\\
&\le C_1\left\{T(a_n)\epsilon_n(x)\left(1-\frac{|x|}{a_n}\right)^{1/2} + \frac{T(a_n)\epsilon_n(x)^{3/2}}{a_n^{1/2}} + T(a_n)\epsilon_n(x)^{3/2}\right\}\\
&\le C_2\left\{T(a_n)\epsilon_n(x)\left(1-\frac{|x|}{a_n}\right)^{1/2} + T(a_n)\epsilon_n(x)^{3/2}\right\}.
\end{aligned}$$

Then

$$nU_n\left(\frac{t}{a_n}\right) \le C_2\left\{nT(a_n)\epsilon_n(x)\left(1-\frac{|x|}{a_n}\right)^{1/2} + nT(a_n)\epsilon_n(x)^{3/2}\right\} \le C_3,$$

by the choice (5.16) of $\epsilon_n(x)$.

Case II. $|a| > 1$. Since $U_n(a) < 0$ (see (2.8)), we have

$$\begin{aligned} U_n\left(\frac{t}{a_n}\right) &= U_n(a+ib) - U_n(a) + U_n(a) \\ &\leq \int_0^1 \log\left[1 + \left(\frac{|b|}{|a|-t}\right)^2\right] \mu_n(t)dt \quad \text{(by (5.6))} \\ &\leq \int_0^1 \log\left[1 + \left(\frac{|b|}{1-t}\right)^2\right] \mu_n(t)dt, \end{aligned}$$

as $|a| - t > 1 - t$, $t \in [0,1]$. The argument of Case I with $a = 1$ then shows that the above integral is bounded above by the right-hand side of (5.21) (with 1 substituted for a there). Thus,

$$\begin{aligned} nU_n\left(\frac{t}{a_n}\right) &\leq nCT(a_n)|b|^{3/2} \int_{-\infty}^{\infty} \log\left[1 + \frac{1}{s^2}\right] |s|^{1/2} ds \\ &\leq C_1 nT(a_n)\epsilon_n(x)^{3/2} \leq C_2. \quad \square \end{aligned}$$

6. Proofs of Theorems 1.2–1.5.

Proof of Theorem 1.2(i). First, let $0 < \beta < 1$, and let, as in (5.1),

$$\epsilon_n := \eta(nT(a_n)^{1/2})^{-1}, \qquad n \geq 1.$$

Then Lemma 2.2 and (5.2) ensure that for $P \in \mathcal{P}_n$, and $|x| \leq a_{\beta n}$,

$$|(Pw)'(x)| \leq \epsilon_n^{-1} e^{\tau_n(x)} \|Pw\|_{L_\infty[-1,1]} C,$$

where, as in (2.13),

$$\tau_n(x) := \begin{cases} Q'(3\epsilon_n)2\epsilon_n, & \text{if } |x| \leq 2\epsilon_n, \\ \left[\frac{Q'(|x|+\epsilon_n)}{|x|} + Q''(|x|+\epsilon_n)\right]2\epsilon_n^2, & \text{if } |x| > 2\epsilon_n. \end{cases}$$

Since $\epsilon_n \to 0$ as $n \to \infty$,

$$Q'(3\epsilon_n)2\epsilon_n \leq C_1, \qquad n \geq 1.$$

Next, from (5.4), $|x| + \epsilon_n < a_n$, if η is small enough. So for $|x| \geq 2\epsilon_n$,

$$Q'(|x| + \epsilon_n)|x|^{-1}\epsilon_n^2 \leq \tfrac{1}{2}Q'(a_n)\epsilon_n \leq CnT(a_n)^{1/2}\epsilon_n \leq C_3,$$

by (3.7) and the choice of ϵ_n. Furthermore, by Lemma 3.1(ii), if $|x| \geq \frac{1}{4}$,

$$Q''(|x| + \epsilon_n)\epsilon_n^2 \leq \frac{a_n}{|x| + \epsilon_n} Q''(a_n)\epsilon_n^2 \leq C_4 T(a_n)^{1/2} n^{-1} \leq C_5,$$

by (3.7), (3.11) and the choice of ϵ_n. If $|x| \leq \frac{1}{4}$,

$$Q''(|x| + \epsilon_n)\epsilon_n^2 \leq (|x| + \epsilon_n)Q''(|x| + \epsilon_n)\epsilon_n \leq \tfrac{1}{2}Q''(1/2)\epsilon_n \leq C_6,$$

by Lemma 3.1(ii). Thus we have shown that

$$\tau_n(x) \le C_7, \qquad |x| \le a_{\beta n}.$$

So for $P \in \mathcal{P}_n$, $n \ge 1$,

$$\|(Pw)'\|_{L_\infty[-a_{\beta n},a_{\beta n}]} \le C_8 n T(a_n)^{1/2} \|Pw\|_{L_\infty[-1,1]}. \tag{6.1}$$

Choosing $\beta = \frac{1}{2}$, and replacing n by $2n$, yields for $P \in \mathcal{P}_n \subset \mathcal{P}_{2n}$,

$$\|(Pw)'\|_{L_\infty[-a_n,a_n]} \le C_8 2n T(a_{2n})^{1/2} \|Pw\|_{L_\infty[-1,1]}. \tag{6.2}$$

Then, as

$$Q'(a_{2n}) = \mathrm{O}(nT(a_{2n})^{1/2}),$$

we obtain

$$\|P'w\|_{L_\infty[-a_n,a_n]} \le C_8 2n T(a_{2n})^{1/2} \|Pw\|_{L_\infty[-1,1]}.$$

The Mhaskar–Saff identity (1.9) then yields Theorem 1.2(i). □

Proof of Theorem 1.2(ii). Suppose that $n \ge 1$, and

$$|x| \le a_n \left(1 - \frac{A}{T(a_{2n})}\right), \tag{6.3}$$

where A is some fixed but large enough positive number. Note that then, by (3.11),

$$|x| \le a_n(1 - 4n^{-2}a_1^{-2}), \tag{6.4}$$

if only A is large enough. Thus (5.7) is satisfied. Let

$$\hat{\epsilon}_n(x) := \eta n^{-1} \left(1 - \frac{|x|}{a_n}\right)^{1/2},$$

where $\eta \in (0,1)$ is fixed and independent of n and x, but small enough. Note then that $\hat{\epsilon}_n(x) \le \epsilon_n(x)$, where $\epsilon_n(x)$ is defined by (5.8). Then Lemma 2.2 and (5.9) ensure that for $P \in \mathcal{P}_n$,

$$|(Pw)'(x)| \le \hat{\epsilon}_n(x)^{-1} e^{\tau_n(x)} \|Pw\|_{L_\infty[-1,1]} C,$$

where

$$\tau_n(x) := \begin{cases} Q'(3\hat{\epsilon}_n(x))2\hat{\epsilon}_n(x), & \text{if } |x| \le 2\hat{\epsilon}_n(x), \\ \left[\frac{Q'(|x|+\hat{\epsilon}_n(x))}{|x|} + Q''(|x| + \hat{\epsilon}_n(x))\right] 2\hat{\epsilon}_n(x)^2, & \text{if } |x| > 2\hat{\epsilon}_n(x). \end{cases}$$

Since $\hat{\epsilon}_n(x) \to 0$ as $n \to \infty$, uniformly for the above range of x, we have

$$\tau_n(x) \le C, \quad \text{if} \quad |x| \le 2\hat{\epsilon}_n(x).$$

Next, if $2\hat{\epsilon}_n(x) \le |x| \le \frac{1}{4}$, then for n large enough,

$$\begin{aligned} \tau_n(x) &= \left[\frac{Q'(|x| + \hat{\epsilon}_n(x))}{|x|} + Q''(|x| + \hat{\epsilon}_n(x))\right] 2\hat{\epsilon}_n(x)^2 \\ &\le 2Q'(|x| + \hat{\epsilon}_n(x))\hat{\epsilon}_n(x) + 2(|x| + \hat{\epsilon}_n(x))Q''(|x| + \hat{\epsilon}_n(x))\hat{\epsilon}_n(x) \\ &\le \frac{2Q'(1/2)\eta}{n} + \frac{2(1/2)Q''(1/2)\eta}{n} \to 0, \end{aligned}$$

as $n \to \infty$, by Lemma 3.1(ii). If $|x| \geq \frac{1}{4}$, then we estimate $\tau_n(x)$ as follows: First note that by (6.4),

$$\frac{\hat{\epsilon}_n(x)}{(1-|x|/a_n)} = \eta n^{-1}\left(1-\frac{|x|}{a_n}\right)^{-1/2} \leq C_1\eta.$$

Hence,

$$\begin{aligned} 1-\frac{|x|+2\hat{\epsilon}_n(x)}{a_n} &= 1-\frac{|x|}{a_n}-\frac{2\hat{\epsilon}_n(x)}{a_n} \\ &\geq \left(1-\frac{|x|}{a_n}\right)\left(1-\frac{2C_1\eta}{a_1}\right) \\ &\geq \frac{1}{2}\left(1-\frac{|x|}{a_n}\right) \geq \frac{\eta}{2}\hat{\epsilon}_n(x), \end{aligned} \tag{6.5}$$

if η is small enough. Then as $|x| \geq \frac{1}{4}$, Lemma 3.1(ii) shows that

$$\begin{aligned} Q''(|x|+\hat{\epsilon}_n(x))\hat{\epsilon}_n(x)^2 &\leq \left(\int_{|x|+\hat{\epsilon}_n(x)}^{|x|+2\hat{\epsilon}_n(x)} Q''(t)dt\right)\frac{|x|+2\hat{\epsilon}_n(x)}{|x|+\hat{\epsilon}_n(x)}\hat{\epsilon}_n(x) \\ &\leq 2\hat{\epsilon}_n(x)Q'(|x|+2\hat{\epsilon}_n(x)) \\ &\leq C_2 n^{-1}\left\{1-\frac{|x|+2\hat{\epsilon}_n(x)}{a_n}\right\}^{1/2} Q'(|x|+2\hat{\epsilon}_n(x)) \\ &\qquad \text{(by (6.5) and the choice of } \hat{\epsilon}_n(x)) \\ &\leq C, \end{aligned}$$

by (3.17) of Lemma 3.3. Also then,

$$Q'(|x|+\hat{\epsilon}_n(x))|x|^{-1}\hat{\epsilon}_n(x)^2 \leq 4Q'(|x|+2\hat{\epsilon}_n(x))\hat{\epsilon}_n(x) \leq C,$$

as above. So

$$\tau_n(x) \leq C,$$

uniformly for $|x| \leq a_n(1-A/T(a_{2n}))$, and we have

$$|(Pw)'(x)| \leq \hat{\epsilon}_n(x)^{-1}C_1\|Pw\|_{L_\infty[-1,1]} \leq \frac{C_2 n}{\sqrt{1-|x|/a_n}}\|Pw\|_{L_\infty[-1,1]}, \tag{6.6}$$

$P \in \mathcal{P}_n$, uniformly for $|x| \leq a_n(1-1/T(a_{2n}))$. On the other hand, if

$$a_n > |x| \geq a_n\left(1-\frac{A}{T(a_{2n})}\right),$$

then (6.2) shows that

$$|(Pw)'(x)| \leq C_3 nT(a_{2n})^{1/2}\|Pw\|_{L_\infty[-1,1]} \leq \frac{C_3 n}{\sqrt{1-|x|/a_n}}\|Pw\|_{L_\infty[-1,1]}.$$

Summarizing, we have shown that

$$\max_{x\in[-a_n,a_n]}\left|(Pw)'(x)\sqrt{1-|x|/a_n}\right| \le C_4 n\|Pw\|_{L_\infty[-1,1]}, \tag{6.7}$$

$P \in \mathcal{P}_n$, $n \ge 1$. Since (3.17) implies that

$$Q'(x)\sqrt{1-|x|/a_n} \le Cn,$$

for $\frac{1}{4} \le |x| \le 1$, and this inequality is trivial for $|x| \le \frac{1}{2}$, we have shown that

$$\max_{x\in[-a_n,a_n]}|P'w|(x)\sqrt{1-|x|/a_n} \le C_4 n\|Pw\|_{L_\infty[-1,1]}, \tag{6.8}$$

$P \in \mathcal{P}_n$, $n \ge 1$, which completes the proof of the theorem. □

Proof of Corollary 1.3. It follows directly from Theorem 1.2(i), (ii), that (1.18) holds for $|x| < a_n$, since

$$\left[\left(1-\frac{|x|}{a_n}\right)^{1/2}+T(a_{2n})^{-1/2}\right]^{-1} \sim \min\left\{\left|1-\frac{|x|}{a_n}\right|^{-1/2}, T(a_{2n})^{1/2}\right\}.$$

We can then reformulate (1.18) for $|x| < a_n$ as

$$\left\{\left(1-\left(\frac{x}{a_n}\right)^2\right)^2+T(a_{2n})^{-2}\right\}[(P'w)(x)]^4 \le Cn^2\|Pw\|^4_{L_\infty[-1,1]}, \tag{6.9}$$

$|x| \le a_n$, $P \in \mathcal{P}_n$. Now $\psi(x) := \{(1-(x/a_n)^2)^2+T(a_{2n})^{-2}\}P'(x)^4 \in \mathcal{P}_{4n}$, and by the Mhaskar–Saff identity applied to $w^4 = e^{-4Q}$, (for which $a_{4n}(4Q) = a_n(Q)$), we have

$$\|\psi w^4\|_{L_\infty[-1,1]} = \|\psi w^4\|_{L_\infty[-a_n,a_n]}.$$

Hence (6.9) holds for $x \in [-1,1]$, and so does (1.18). □

Proof of Corollary 1.4(i). It obviously suffices to estimate $T(a_n)$ for $W_{0,\alpha} = e^{-Q}$ given by (1.4). A straightforward calculation shows that

$$T(x) = \frac{2(1+\alpha x^2)}{1-x^2} \sim (1-|x|)^{-1}, \qquad x \in (-1,1).$$

Furthermore,

$$Q'(x) \sim x(1-|x|)^{-\alpha-1}, \qquad x \in (-1,1),$$

and

$$T(x) \sim \frac{Q'(x)}{Q(x)}, \qquad x \text{ near } 1,$$

so by (3.9), for large enough n, $Q'(a_n) \sim nT(a_n)^{1/2}$, which implies

$$(1-a_n)^{-\alpha-1} \sim n(1-a_n)^{-1/2},$$

and hence

$$1-a_n \sim n^{-1/(\alpha+1/2)}.$$

Then $T(a_n) \sim n^{1/(\alpha+1/2)}$, and Theorem 1.2(i) yields Corollary 1.4(i). □

Proof of Corollary 1.4(ii). Let

$$\exp_0(x) := x, \quad \exp_\ell(x) := \exp(\exp_{\ell-1}(x)), \quad \ell \geq 1,$$

and

$$F_0(x) := 1, \quad F_\ell(x) := \prod_{j=1}^{\ell} \exp_j(x), \quad \ell \geq 1,$$

and

$$Q(x) := \exp_k((1-x^2)^{-\alpha}).$$

Note that

$$\frac{d}{dx}\exp_\ell(x) = F_\ell(x), \qquad \ell \geq 0,$$

and

$$\frac{d}{dx}F_\ell(x) = F_\ell(x)\sum_{j=0}^{\ell-1} F_j(x).$$

A straightforward calculation shows that

$$Q'(x) = F_k((1-x^2)^{-\alpha})2\alpha x(1-x^2)^{-\alpha-1},$$

and

$$T(x) = \frac{2}{1-x^2}\left\{\sum_{\ell=0}^{k-1} F_\ell((1-x^2)^{-\alpha})\alpha x^2(1-x^2)^{-\alpha} + 1 + \alpha x^2\right\}.$$

Hence for x close to 1,

$$Q'(x) \sim F_k((1-x^2)^{-\alpha})(1-x^2)^{-\alpha-1},$$
$$T(x) \sim F_{k-1}((1-x^2)^{-\alpha})(1-x^2)^{-\alpha-1},$$

and

$$T(x) \sim \frac{Q'(x)}{Q(x)}.$$

Then by (3.9), $Q'(a_n) \sim nT(a_n)^{1/2}$ implies

$$F_k((1-a_n^2)^{-\alpha})(1-a_n^2)^{-\alpha-1} \sim n\left[F_{k-1}((1-a_n^2)^{-\alpha})\right]^{1/2}(1-a_n^2)^{-(\alpha+1)/2}.$$

Taking logarithms shows that for n large enough,

$$\exp_{k-1}((1-a_n^2)^{-\alpha}) = \log n - \frac{1}{2}\sum_{j=0}^{k-2}\exp_j((1-a_n^2)^{-\alpha}) - \frac{\alpha+1}{2}\log(1-a_n^2) + \mathrm{O}(1),$$

which implies that as $n \to \infty$,

$$\exp_{k-1}((1-a_n^2)^{-\alpha}) = \log n + O\left(\frac{\log\log n}{\log n}\right).$$

From this it readily follows that for $0 \le j \le k-1$, and n large enough,

$$\exp_j((1-a_n^2)^{-\alpha}) = \log_{k-j} n + o(1).$$

Then for n large enough,

$$\begin{aligned} T(a_n) \sim F_{k-1}((1-a_n^2)^{-\alpha})(1-a_n^2)^{-\alpha-1} &\sim \left(\prod_{j=1}^{k-1} \log_{k-j} n\right) (\log_k n)^{(\alpha+1)/\alpha} \\ &= \left(\prod_{j=1}^{k-1} \log_j n\right) (\log_k n)^{(\alpha+1)/\alpha}. \end{aligned}$$

Now Theorem 1.2(i) yields the result. □

Proof of Corollary 1.4(iii). Here $w = e^{-Q}$, where $Q(x) = -\alpha \log(1-x^2)$. Then

$$Q'(x) = \frac{2\alpha x}{1-x^2},$$

and a simple calculation shows that

$$T(x) = \frac{2}{1-x^2}.$$

Then (1.8) yields

$$\begin{aligned} n &\sim \int_{1/2}^1 \frac{Q'(a_n t)}{\sqrt{1-t^2}} dt \sim \int_{1/2}^1 \frac{1}{\sqrt{1-t}(1-a_n t)} dt \\ &= \int_0^{1/2} \frac{ds}{\sqrt{s}(1-a_n+a_n s)} \sim \int_0^{1-a_n} \frac{ds}{\sqrt{s}(1-a_n)} + \int_{1-a_n}^{1/2} \frac{ds}{s} \\ &\sim (1-a_n)^{-1/2} + \log\left[\frac{1}{2(1-a_n)}\right]. \end{aligned}$$

Then we deduce that

$$1 - a_n \sim n^{-2},$$

and hence

$$T(a_n) \sim n^2.$$

Again, Theorem 1.2(i) yields the result. □

Proof of Theorem 1.5(a). From (6.2) and (6.7),

$$|(Pw)'(x)| \le C_5 n \min\{T(a_{2n})^{1/2}, (1-|x|/a_n)^{-1/2}\} \|Pw\|_{L_\infty[-1,1]}$$

for $|x| < a_n$ and $P \in \mathcal{P}_n$. This immediately yields (1.22). □

Proof of Theorem 1.5(b). We remark that (1.23) is implied by (1.22), in a somewhat stronger form, if $|x| \le \frac{1}{2}$. So we may assume that $|x| \ge \frac{1}{2}$. Let

$$\epsilon_n(x) := \eta \min\left\{(nT(a_{2n}))^{-2/3}, \left(nT(a_{2n})\sqrt{1-|x|/a_n}\right)^{-1}\right\}.$$

Then by Lemma 2.2 and Theorem 5.3,

$$|(Pw)'(x)| \le \epsilon_n(x)^{-1} e^{\tau_n(x)} \|Pw\|_{L_\infty[-1,1]} C,$$

$P \in \mathcal{P}_n$, $\frac{1}{2} \le |x| \le a_n$, where

$$\begin{aligned}\tau_n(x) &= \left[\frac{Q'(|x|+\epsilon_n(x))}{|x|} + Q''(|x|+\epsilon_n(x))\right] 2\epsilon_n(x)^2 \\ &\le 4[Q'(|x|+\epsilon_n(x)) + Q''(|x|+\epsilon_n(x))]\epsilon_n(x)^2.\end{aligned}$$

Now, by (3.11),

$$\epsilon_n(x) \le \eta(nT(a_{2n}))^{-2/3} \le \frac{C\eta}{T(a_{2n})}.$$

Then by (3.13),

$$|x| + \epsilon_n(x) \le a_n\left(1 + \frac{C_1\eta}{T(a_{2n})}\right) \le a_{2n},$$

if only η is small enough. Hence the monotonicity of $Q'(u)$ and $uQ''(u)$ yield

$$\begin{aligned}\tau_n(x) &\le C[Q'(a_{2n}) + Q''(a_{2n})]\epsilon_n^2(x) \\ &\le C_1 nT(a_{2n})^{3/2}(nT(a_{2n}))^{-4/3} \quad \text{(by (3.7) and the choice of } \epsilon_n(x)) \\ &= C_1 n^{-1/3} T(a_{2n})^{1/6} \le C_2,\end{aligned}$$

by (3.11). Hence for $|x| < a_n$ and $P \in \mathcal{P}_n$,

$$\begin{aligned}|(Pw)'(x)| &\le \epsilon_n(x)^{-1} C_3 \|Pw\|_{L_\infty[-1,1]} \\ &\le C_4 \max\left\{(nT(a_{2n}))^{2/3}, nT(a_{2n})\left(1-\frac{|x|}{a_n}\right)^{1/2}\right\} \|Pw\|_{L_\infty[-1,1]}.\end{aligned}$$

Then (1.23) follows for $|x| < a_n$ and (1.24) follows by continuity. □

7. Proof of Theorem 1.6 and Corollary 1.7.

Proof of Theorem 1.6. Let $P \in \mathcal{P}_{n-1}$ and choose $\xi \in [-a_n, a_n]$ such that

$$|Pw|(\xi) = \|Pw\|_{L_\infty[-1,1]}.$$

Let

$$\delta := \tfrac{1}{2}(CnT(a_{2n-2})^{1/2})^{-1}, \tag{7.1}$$

where C is as in (1.22). Then for $t \in (\xi - \delta, \xi + \delta) \cap (-1,1)$, there is a $\zeta \in (\xi - \delta, \xi + \delta) \cap (-1,1)$ such that

$$\begin{aligned}|Pw|(t) &= |(Pw)(\xi) + (Pw)'(\zeta)(t-\xi)| \\ &\geq |Pw|(\xi) - \|(Pw)'\|_{L_\infty[-a_n,a_n]}\delta \\ &\geq \|Pw\|_{L_\infty[-1,1]} - \left(\frac{\delta^{-1}}{2}\right)\|Pw\|_{L_\infty[-1,1]}\delta = \frac{1}{2}\|Pw\|_{L_\infty[-1,1]},\end{aligned}$$

by (1.22). So if $x \in (-1,1)$,

$$\int_{-1}^{1} \frac{(Pw)^2(t)dt}{(Pw)^2(x)} \geq \int_{(\xi-\delta,\xi+\delta)\cap(-1,1)} \frac{1}{4}\|Pw\|^2_{L_\infty[-1,1]}dt/(Pw)^2(x) \geq \frac{\delta}{4}.$$

Thus

$$\lambda_n(w^2,x)w^{-2}(x) = \inf_{P\in\mathcal{P}_{n-1}} \int_{-1}^{1}(Pw)^2(t)dt/(Pw)^2(x) \geq \frac{\delta}{4}.$$

Taking account of the definition (7.1) of δ, we obtain (1.25). □

Proof of Corollary 1.7. We use a very standard argument (see, for example, [17]) but provide the details for the reader's convenience.

Step 1. Let

$$\rho_n := nT(a_{2n-2})^{1/2}, \qquad n \geq 1.$$

By Theorem 1.6, in the form proved above,

$$(Pw)^2(x) \leq C\rho_n \int_{-1}^{1}(Pw)^2(t)dt \quad \forall x \in [-1,1], \quad P \in \mathcal{P}_{n-1}.$$

Hence

$$\|Pw\|_{L_\infty[-1,1]} \leq C\rho_n^{1/2}\|Pw\|_{L_2[-1,1]}.$$

Applying this to $w^k = e^{-kQ}$, and noting that $a_{2kn}(kQ) = a_{2n}(Q)$ yields for $P \in \mathcal{P}_{n-1}$,

$$\|P^k w^k\|_{L_\infty[-1,1]} \leq C_1\rho_n^{1/2}\|P^k w^k\|_{L_2[-1,1]}, \tag{7.2}$$

and hence

$$\|Pw\|_{L_\infty[-1,1]} \leq C\rho_n^{1/(2k)}\|Pw\|_{L_{2k}[-1,1]}.$$

So we have (1.27) for $q = \infty$ and $p = 2k$.

Step 2. Let $p > 0$, and choose an integer k such that $2k > p$. Then by (7.2),

$$\begin{aligned}\|Pw\|^{2k}_{L_\infty[-1,1]} &\leq C\rho_n \int_{-1}^{1}(Pw)^{2k}(t)dt \\ &\leq C\rho_n\|Pw\|^{2k-p}_{L_\infty[-1,1]}\int_{-1}^{1}|Pw|^p(t)dt;\end{aligned}$$

so if P is not identically zero,

$$\|Pw\|^p_{L_\infty[-1,1]} \le C\rho_n \int_{-1}^1 |Pw|^p(t)dt.$$

This is still trivially true if P is identically zero. Then (1.27) follows for $q = \infty$ and any $p > 0$.

Step 3. We let $p > 0$, and may assume $q < \infty$. Then

$$\begin{aligned}\|Pw\|^q_{L_q[-1,1]} &= \int_{-1}^1 |Pw|^{q-p}(t)|Pw|^p(t)dt \\ &\le \|Pw\|^{q-p}_{L_\infty[-1,1]}\|Pw\|^p_{L_p[-1,1]} \\ &\le (C\rho_n)^{(q-p)/q}\|Pw\|^{q-p}_{L_q[-1,1]}\|Pw\|^p_{L_p[-1,1]},\end{aligned}$$

by (1.27) for the case already proven; so if P is not identically zero,

$$\|Pw\|^p_{L_q[-1,1]} \le (C\rho_n)^{(q-p)/q}\|Pw\|^p_{L_p[-1,1]}. \qquad \square$$

REFERENCES

[1] Z. Ditzian and V. Totik, *Moduli of Smoothness*, Springer Ser. Comput. Math., 9, Springer-Verlag, Berlin, 1987.

[2] T. Erdelyi, *Nikolskii-type inequalities for generalized polynomials and zeros of orthogonal polynomials*, J. Approx. Theory, 66 (1991), pp. 80–92.

[3] T. Erdelyi, A. Máté, and P. Nevai, *Inequalities for generalized polynomials*, Constr. Approx., 8 (1992), pp. 241–255.

[4] A. L. Levin and D. S. Lubinsky, *L_∞ Markov and Bernstein inequalities for Freud weights*, SIAM J. Math. Anal., 21 (1990), pp. 1065–1082.

[5] D.S. Lubinsky, *Strong asymptotics for extremal errors and polynomials associated with Erdös-type weights*, Pitman Res. Notes Math., 202, Longmans, Harlow, 1989.

[6] ———, *L_∞ Markov and Bernstein inequalities for Erdös weights*, J. Approx. Theory, 60 (1991), pp. 188–230.

[7] ———, *Hermite and Hermite-Fejer interpolation and associated product integration rules on the real line: the L_∞ theory*, J. Approx. Theory, to appear.

[8] D. S. Lubinsky and T. Z. Mthembu, *The supremum norm of reciprocals of Christoffel functions for Erdös weights*, J. Approx. Theory, 63 (1990), pp. 255–266.

[9] ———, *L_p Markov and Bernstein inequalities for Erdös weights*, J. Approx. Theory, 65 (1991), pp. 301–321.

[10] D. S. Lubinsky and P. Nevai, *Markov–Bernstein inequalities revisited*, J. Approx. Theory Appl., 3 (1987), pp. 98–119.

[11] D. S. Lubinsky and E. B. Saff, *Strong asymptotics for extremal polynomials associated with weights on* **R**, Lecture Notes in Math., 1305, Springer-Verlag, Berlin, 1988.

[12] ———, *Asymptotics for non-Szegő weights on* $[-1,1]$, in Approximation Theory VI, Vol. II, C. K. Chui, L. L. Schumaker, J. D. Ward, eds., Academic Press, San Diego, 1989, pp. 409–412.

[13] H. N. Mhaskar and E. B. Saff, *Where does the sup-norm of a weighted polynomial live?*, Constr. Approx., 1 (1985), pp. 71–91.

[14] ———, *Where does the L_p-norm of a weighted polynomial live?*, Trans. Amer. Math. Soc., 303 (1987), pp. 109–124; Errata, 308 (1988), p. 431.

[15] P. Nevai, *Orthogonal polynomials*, Mem. Amer. Math. Soc., 18 (1979).

[16] P. Nevai and G. Freud, *Orthogonal polynomials and Christoffel functions: a case study*, J. Approx. Theory, 48 (1986), pp. 3–167.

[17] P. Nevai and V. Totik, *Weighted polynomial inequalities*, Constr. Approx., 2 (1986), pp. 113–127.

[18] ———, *Sharp Nikolskii inequalities with exponential weights*, Anal. Math., 13 (1987), pp. 261–267.

SIAM J. MATH. ANAL.
Vol. 24, No. 3, pp. 557–570, May 1993

001

UNIVERSAL BOUNDS FOR THE LOW EIGENVALUES OF NEUMANN LAPLACIANS IN N DIMENSIONS*

MARK S. ASHBAUGH† AND RAFAEL D. BENGURIA‡

Abstract. The authors consider bounds on the Neumann eigenvalues of the Laplacian on domains in $\mathbb{R}^n$ in the light of their recent results on Dirichlet eigenvalues, in particular, their proof of the Payne–Pólya–Weinberger conjecture via spherical rearrangement. They prove the bound $1/\mu_1 + 1/\mu_2 \geq A/2\pi$ for the first two nonzero Neumann eigenvalues for an arbitrary bounded domain Ω in two dimensions and also the stronger (and optimal) bound $\mu_2 \leq \pi(j'_{1,1})^2/A$ for domains having a 4-fold rotational symmetry. (Here $j'_{1,1} \approx 1.84118$ denotes the first positive zero of the derivative of the Bessel function $J_1(x)$ and A is the area of the domain Ω.) The authors also obtain analogues of these results for domains in $\mathbb{R}^n$. Previous results in this vein are due to Szegö, who proved $\mu_1 \leq \pi(j'_{1,1})^2/A$ and $1/\mu_1 + 1/\mu_2 \geq 2A/\pi(j'_{1,1})^2$ for simply connected domains in $\mathbb{R}^2$, and to Weinberger, who proved the general result $\mu_1 \leq (C_n/|\Omega|)^{2/n} p^2_{n/2,1}$ for arbitrary domains in $\mathbb{R}^n$ (here $C_n = \pi^{n/2}/\Gamma(n/2+1) =$ volume of the unit ball in $\mathbb{R}^n$, and $p_{\nu,k}$ denotes the kth positive zero of the derivative of $x^{1-\nu}J_\nu(x)$, where $J_\nu(x)$ represents the standard Bessel function of the first kind of order ν).

Key words. eigenvalues of Neumann Laplacians, universal eigenvalue inequalities, zeros of Bessel functions

AMS subject classifications. 35P15, 49Gxx, 35J05, 33A40

1. Introduction.

In this paper we consider bounds on combinations of the low eigenvalues of the Neumann problem

$$-\Delta u = \mu u \quad \text{on } \Omega, \tag{1.1a}$$

$$\frac{\partial u}{\partial n} = 0 \quad \text{on } \partial\Omega \tag{1.1b}$$

for Ω a bounded domain in $\mathbb{R}^n$ having a smooth boundary. It is well known that this problem has spectrum $\{\mu_i\}_{i=0}^{\infty}$ diverging to infinity and satisfying

$$0 = \mu_0 < \mu_1 \leq \mu_2 \leq \cdots . \tag{1.2}$$

For the closely related Dirichlet problem, with eigenvalues

$$0 < \lambda_1 < \lambda_2 \leq \lambda_3 \leq \cdots \longrightarrow \infty, \tag{1.3}$$

we now have a variety of universal bounds on eigenvalues, such as

$$\lambda_{m+1} - \lambda_m \leq \frac{4}{mn}(\lambda_1 + \cdots + \lambda_m), \tag{1.4}$$

*Received by the editors August 14, 1991; accepted by the editors (in revised form) September 3, 1992.

†Department of Mathematics, University of Missouri, Columbia, Missouri 65211. This research was partially supported by National Science Foundation grants DMS-9114162 and INT-9123481 (mathmsa@mizzou1.missouri.edu).

‡Facultad de Física, P. Universidad Católica de Chile, Casilla 306, Santiago 22, Chile. This author's research was supported in part by Fondo Nacional de Ciencia y Tecnología (Chile) project 1238-90 (rbenguri@lascar.puc.cl).

where n = spatial dimension (i.e., for $\Omega \subset I\!\!R^n$) due to Payne, Pólya, and Weinberger [34], [35] (and to Thompson [43] for the relatively straightforward extension from 2 to n dimensions),

$$\lambda_3 \leq 5\lambda_1 + \frac{\lambda_1^2}{\lambda_2} - \lambda_2 \tag{1.5}$$

for $n = 2$ due to Brands [11],

$$\frac{\lambda_3}{\lambda_1} \leq \frac{7(15+\sqrt{345})}{60} \approx 3.9170 \tag{1.6}$$

and

$$\frac{\lambda_2 + \lambda_3}{\lambda_1} \leq \frac{15+\sqrt{345}}{6} \approx 5.5957 \tag{1.7}$$

for $n = 2$ due to Marcellini [28], and

$$\frac{\lambda_2}{\lambda_1} \leq \frac{j_{n/2,1}^2}{j_{n/2-1,1}^2} = \left.\frac{\lambda_2}{\lambda_1}\right|_{n\text{-dimensional ball}}, \tag{1.8}$$

$$\frac{\lambda_4}{\lambda_2} < \frac{j_{n/2,1}^2}{j_{n/2-1,1}^2}, \tag{1.9}$$

and

$$\sum_{k=1}^{n} \frac{1}{\lambda_{k+1} - \lambda_1} \geq \frac{2j_{n/2-1,1}^2 + n(n-4)}{6\lambda_1} \tag{1.10}$$

due to us [3]–[7]. (Here and throughout the paper $j_{\nu,k}$ denotes the kth positive zero of the Bessel function $J_\nu(x)$.) Other contributors to these developments include Hile and Protter [20] and Chiti [14], [15]; indeed, (1.10) was developed directly from results of Chiti in [14] and [15] while (1.8) and (1.9) were developed using ideas introduced by Chiti in [15].

For a given dimension all the inequalities given above exhibit a universal bound on combinations of eigenvalues in the sense that the bound holds without regard to the particular size or shape of the domain.

In his surveys of this subject (which also discuss similar results for certain operators involving powers of the Laplacian and in particular the biharmonic operator), Protter remarked that no bound for the Neumann eigenvalues similar to (1.4) had yet been found [39, pp. 192–193] and [40, p. 119]. Indeed, almost simultaneously it was shown that no such bound for Neumann eigenvalues could possibly hold. In [16] (see also Shubin's comments in [2]) Colin de Verdière showed that given any nondecreasing sequence of positive numbers of length ℓ there exists a domain Ω having these values as $\mu_1, \mu_2, \ldots, \mu_\ell$. Thus $\mu_1, \ldots, \mu_\ell$ are completely unconstrained, aside from the obvious $0 < \mu_1 \leq \mu_2 \leq \cdots \leq \mu_\ell$ for any finite ℓ. While this result might suggest that it is pointless to pursue universal eigenvalue inequalities for the Neumann problem, this is really only partly true. For example, if in two dimensions one adds the area A (volume

$|\Omega|$ in dimension n) to the list of objects between which one seeks constraints then some interesting inequalities may be obtained. Returning to the Dirichlet problem for a moment, there is the celebrated Faber–Krahn inequality [17], [23], [24],

$$\lambda_1 \geq \frac{\pi j_{0,1}^2}{A}, \tag{1.11}$$

which says that $\lambda_1 A$ takes its minimum when Ω is a disk and, similarly, in n dimensions we have

$$\lambda_1 \geq \left(\frac{C_n}{|\Omega|}\right)^{2/n} j_{n/2-1,1}^2, \tag{1.12}$$

where C_n = volume of the unit ball in $I\!R^n$ and again $j_{\nu,k}$ denotes the kth positive zero of the Bessel function $J_\nu(x)$. We also have

$$\lambda_2 > \frac{2\pi j_{0,1}^2}{A} \tag{1.13}$$

in $I\!R^2$ and an analogous result in $I\!R^n$ (due to Peter Szego; see [38, p. 336]).

There are similar results for the Neumann problem. In two dimensions we have the result

$$\mu_1 \leq \frac{\pi p_{1,1}^2}{A} \tag{1.14}$$

of Szegö [42], which he proved for simply connected domains via conformal mapping techniques ("the method of conformal transplantation"; see Bandle [10], for example). Later, Weinberger [44], using more general methods, showed that (1.14) and its n-dimensional analogue,

$$\mu_1 \leq \left(\frac{C_n}{|\Omega|}\right)^{2/n} p_{n/2,1}^2, \tag{1.15}$$

hold for arbitrary domains in $I\!R^2$ and $I\!R^n$, respectively. Here $p_{\nu,k}$ denotes the kth positive zero of the derivative of $x^{1-\nu}J_\nu(x)$. Also Szegö and Weinberger noticed that Szegö's proof of (1.14) for simply connected domains in $I\!R^2$ extends to prove the bound

$$\frac{1}{\mu_1} + \frac{1}{\mu_2} \geq \frac{2A}{\pi p_{1,1}^2} \tag{1.16}$$

for such domains. The bounds of Szegö and Weinberger are optimal (isoperimetric) with equality if and only if Ω is a disk (n-dimensional ball in the case of Weinberger's result (1.15)). It might also be mentioned that Szegö's work on the maximum of $A\mu_1$ was motivated by an earlier conjecture and partial proof of the result (1.14) by Kornhauser and Stakgold [22], who were interested in proving that $\mu_1 < \lambda_1$ in two dimensions and who proposed the route $\mu_1(\Omega) \leq \mu_1(D) < \lambda_1(D) \leq \lambda_1(\Omega)$, where D denotes the disk of the same area as Ω. Pólya [37] then proved $\mu_1 < \lambda_1$ by other means, which actually showed that $\mu_1 \leq 4\pi/A$ and hence that $\mu_1/\lambda_1 \leq 4/j_{0,1}^2 \approx .6917$. With Szegö's proof of (1.14) the constant here is reduced to its optimum value, i.e., $\mu_1/\lambda_1 \leq p_{1,1}^2/j_{0,1}^2 \approx .5862$ with equality if and only if Ω is a disk, at least for the case of simply connected domains (in $\mathbb{R}^2$). Weinberger's subsequent generalization (1.15) then widened the class of domains to which this optimal inequality applies as well as making possible the extension to n dimensions, $\mu_1/\lambda_1 \leq p_{n/2,1}^2/j_{n/2-1,1}^2$. Of course, all of these results make use of the Faber–Krahn inequality, (1.11) or (1.12). Further results along the lines of the inequality $\mu_1 < \lambda_1$ suggested by Kornhauser and Stakgold are due to Payne [32], who proved $\mu_{k+1} < \lambda_k$ $(k = 1, 2, \dots)$ for a convex domain in $\mathbb{R}^2$ with a C^2 boundary, and to Levine and Weinberger [26] (see also [25]), who proved various elegant generalizations of this result to $\mathbb{R}^n$, including the result $\mu_{k+n-1} < \lambda_k$ $(k = 1, 2, \dots)$ for a bounded convex domain in $\mathbb{R}^n$ $(n \geq 2)$ with a sufficiently smooth boundary. At about the same time Aviles [9] showed that $\mu_k < \lambda_k$ $(k = 1, 2, \dots)$ for a bounded domain in $\mathbb{R}^n$ $(n \geq 2)$ with smooth boundary if the mean curvature of the boundary is nonnegative at each of its points. Aviles' result also follows from the general approach of Levine and Weinberger. Of course, the result of Aviles only becomes of interest when $n \geq 3$ since in two dimensions nonnegative mean curvature of the boundary at each point is equivalent to convexity of the domain, and Payne's result $\mu_{k+1} < \lambda_k$ (the $n = 2$ case of the Levine–Weinberger result stated above) is always as strong or stronger than Aviles' $\mu_k < \lambda_k$. Payne has suggested that perhaps $\mu_k \leq \lambda_k$ $(k = 1, 2, \dots)$ holds for any bounded domain in $\mathbb{R}^n$ $(n \geq 2)$ [25] (the $n = 1$ case of this holds as an equality). This conjecture of Payne was recently proved by Friedlander [18] under the hypothesis of a C^1 boundary (see also the related paper of Mazzeo [30]). It might also be noted that such a result cannot hold in general for bounded domains on manifolds; see [12, p. 83] or [13, p. 44, Thm. 3]. Other interesting inequalities between Dirichlet and Neumann eigenvalues have been obtained by Hersch and others; see Hersch's comments on Pólya's paper [37] which appear on pp. 509–510 of [19] for references to these.

In this paper we attempt to go beyond (1.16) in various ways.

Our results are more modest than (1.16) in that they are not optimal but on the other hand they generalize to any dimension n and they are not restricted to simply connected domains. What we prove is that

$$(1.17) \qquad \frac{1}{\mu_1} + \frac{1}{\mu_2} \geq \frac{A}{2\pi}$$

(observe that $p_{1,1} = j_{1,1}' \approx 1.84118 < 2$) for arbitrary domains in two dimensions (this inequality is a special case of a result of Bandle [10, p. 127] which applies to membranes of variable density, but her method of proof, which is based on Szegö's, requires that Ω be simply connected and does not generalize to the n-dimensional case) and its generalization to n dimensions,

$$(1.18) \qquad \frac{1}{\mu_1} + \frac{1}{\mu_2} + \cdots + \frac{1}{\mu_n} \geq \frac{n}{n+2}\left(\frac{|\Omega|}{C_n}\right)^{2/n}.$$

We also prove the optimal bound

$$(1.19) \qquad \mu_2 \leq \frac{\pi p_{1,1}^2}{A}$$

for domains in $\mathbb{R}^2$ with 4-fold rotational symmetry (with equality if and only if the domain is a disk) and a generalization of this to n dimensions. Finally, we investigate the tightness of the bound (1.18) asymptotically as n goes to infinity. We show that

(1.18) is relatively tight in the sense that the first term of the asymptotic expansion of the right-hand side agrees with that of the quantity $1/\mu_1 + 1/\mu_2 + \cdots + 1/\mu_n$ when Ω is a ball (this value is $(n/p_{n/2,1}^2)\left(|\Omega|/C_n\right)^{2/n}$). Similarly, the bound

$$\mu_1 \le (n+2)\left(\frac{C_n}{|\Omega|}\right)^{2/n}, \tag{1.20}$$

which follows from (1.18), is seen to be a relatively tight approximation to the optimal bound (1.15) of Weinberger. These asymptotic results follow from inequalities for the zeros $p_{\nu,1}$ which were recently derived by Lorch and Szego [27]. While it seems likely that these zeros might have been investigated previously, we were unable to find any such information in the literature (see related comments below concerning the Neumann eigenvalues of the unit ball in n dimensions).

Since our results are not optimal (aside from (1.19) and its n-dimensional generalizations) the possibility is left open that the bound

$$\frac{1}{\mu_1} + \frac{1}{\mu_2} + \cdots + \frac{1}{\mu_n} \ge \frac{n(|\Omega|/C_n)^{2/n}}{p_{n/2,1}^2} \tag{1.21}$$

might hold. This bound would be isoperimetric with equality if and only if Ω were an n-dimensional ball (presumably) and would imply Weinberger's bound (1.15).

There is a loose analogy between the ratio results for Dirichlet eigenvalues with which we started our discussion and the results for Neumann eigenvalues discussed above. In particular, if we identify μ_k with $\lambda_{k+1} - \lambda_1$ for $k = 1, 2, 3, \ldots$ and then identify $|\Omega|^{-2/n}$ with λ_1 (or vice versa), then our bound (1.8) for λ_2/λ_1 is identified with Weinberger's bound (1.15) for $|\Omega|^{2/n}\mu_1$ in the sense that these upper bounds are in each case optimal with the maximum being taken on if and only if Ω is an n-dimensional ball. The methods of proof are even almost identical, the main difference being that since in the Neumann case $\mu_0 = 0$ and the corresponding eigenfunction is constant, Weinberger's proof is somewhat simpler (in particular, he is not faced with proving properties of ratios of Bessel functions). Similarly, the bounds (1.17) and (1.18) that we establish in this paper can be viewed as the Neumann analogues of our extension (1.10) of certain inequalities of Chiti. Finally, we mention the further conjecture (for $\Omega \subset \mathbb{R}^2$)

$$\frac{\lambda_2 + \lambda_3}{\lambda_1} \le \frac{2j_{1,1}^2}{j_{0,1}^2} \approx 5.077 \tag{1.22}$$

of Payne, Pólya, and Weinberger [35] for the Dirichlet problem and its connection to Szegö's bound (1.16) and its possible generalization (1.21). The n-dimensional generalization of the conjecture (1.22) would seem to be

$$\frac{\lambda_2 + \cdots + \lambda_{n+1}}{\lambda_1} \le \frac{nj_{n/2,1}^2}{j_{n/2-1,1}^2} \tag{1.23}$$

(also only a conjecture). However, at this point the analogy appears to break down. The Neumann analogue of (1.22)

$$A(\mu_1 + \mu_2) \le 2\pi p_{1,1}^2 \tag{1.24}$$

is definitely *not* true. It does not even hold for all rectangles, whereas (1.22) does. This explains why we look at the quantity $(1/\mu_1 + \cdots + 1/\mu_n)|\Omega|^{-2/n}$ and not at $|\Omega|^{2/n}(\mu_1 + \cdots + \mu_n)$ in this paper. Going the other way, the Dirichlet analogue of the conjecture (1.21) is

$$\sum_{k=1}^{n} \frac{1}{\lambda_{k+1} - \lambda_1} \geq \frac{n}{\left[j_{n/2,1}^2 / j_{n/2-1,1}^2 - 1\right] \lambda_1}, \tag{1.25}$$

which would follow from the conjecture (1.23). While we do not prove (1.21) here, the bounds (1.17) and (1.18) that we do establish are weaker versions of it in the same way that the bound (1.10) is a slightly weaker inequality than the conjectured bound (1.25) above. In both cases the closeness of the approximation for large n is borne out by the asymptotics as n goes to infinity.

One last remark seems in order before we close our introduction. This concerns the Neumann eigenvalues of the unit ball in n dimensions ($n \geq 2$). Somewhat surprisingly, these eigenvalues seem never to have been studied systematically. Since these eigenvalues are determined by the zeros $\tilde{p}_{\nu+1,k}^{(\ell)}$ of the functions $[x^{-\nu} J_{\nu+\ell}(x)]'$ for $\nu = n/2 - 1$ and $\ell = 0, 1, 2, \ldots$ some of them are readily found using known results. In particular, those for the two-dimensional case, for which $\nu = 0$, are all to be found from the roots $j'_{\ell,k}$ of the functions $J'_\ell(x)$ for $\ell = 0, 1, 2, \ldots$. Similarly, for the three-dimensional case the eigenvalues may be found in terms of the roots $a'_{\ell,k}$ of the derivatives of the spherical Bessel functions $j_\ell(x)$ for $\ell = 0, 1, 2, \ldots$. (We follow Abramowitz and Stegun [1] for all our notation concerning Bessel, and related functions; in particular, spherical Bessel functions are defined in eqn. 101.1 on p. 437.) Beyond these cases, only that of spherically symmetric eigenfunctions and their associated eigenvalues can be treated in terms of tabulated results. Since in that case $\ell = 0$ and since $x^\nu \left[x^{-\nu} J_\nu(x)\right]' = -J_{\nu+1}(x)$ the radial eigenvalues are easily found from the zeros $j_{n/2,k}$ of $J_{n/2}(x)$. However, in no other cases beyond $n = 3$ do the Neumann eigenvalues of the n-dimensional ball seem to be readily computable in terms of tabulated quantities. The functions $r^{-\nu} J_{\nu+\ell}(r)$, where $\nu = n/2 - 1$ and $\ell = 0, 1, 2, \ldots$ (and perhaps including an additional constant factor) might well be called "ultraspherical Bessel functions" (as in ultraspherical polynomials) in analogy with the three-dimensional case, since they arise naturally when solving the Helmholtz equation by separation of variables on a ball in $I\!R^n$. The parameter ℓ indexes the eigenvalues of the angular part of the Laplacian in spherical coordinates (these eigenvalues are given by $\ell(\ell + n - 2)$ in dimension n). In particular, for the n-dimensional unit ball $\mu_1 = \mu_2 = \cdots = \mu_n$ occur as $\ell = 1$ eigenvalues and would be denoted $(\tilde{p}_{n/2,1}^{(1)})^2$ in the notation above. However, since only the $\ell = 1$ case ever occurs in this paper, and since the Bessel function then occurring in $[x^{-\nu} J_{\nu+\ell}(x)]'$ has order $\nu + 1 = n/2$, it is convenient for us to use the notation $p_{n/2,1}$, rather than $\tilde{p}_{n/2,1}^{(1)}$, to denote the relevant zeros.

2. Universal bounds for sums of reciprocals of Neumann eigenvalues.

In this section we prove the inequalities (1.17) and (1.18). Specifically, we prove the following.

THEOREM 2.1. *Consider the Neumann problem* (1.1) *for a bounded domain* $\Omega \subset I\!R^n$ ($n \geq 2$) *having smooth boundary. Then the first n nonzero eigenvalues satisfy*

$$\frac{1}{\mu_1} + \frac{1}{\mu_2} + \cdots + \frac{1}{\mu_n} \geq \frac{n}{n+2} \left(\frac{|\Omega|}{C_n} \right)^{2/n}, \tag{2.1}$$

where $|\Omega|$ = *Lebesgue measure of* Ω *and* $C_n = \pi^{n/2}/\Gamma(n/2+1)$ = *the volume of the unit ball in* $I\!R^n$.

Proof. The proof is a relatively easy application of the variational characterization of the eigenvalues μ_i. To be specific, we note that if we let $\{u_i\}_{i=0}^{\infty}$ be an orthonormal set of eigenfunctions for (1.1) in correspondence with the μ_i (so that $-\Delta u_i = \mu_i u_i$ for $i = 0, 1, 2, \ldots$) and we let $\mathcal{H}$ denote the Sobolev space $H^1(\Omega)$, then the eigenvalues μ_i can be characterized by

$$\mu_i = \min_{\substack{u \in \mathcal{H}\setminus\{0\} \\ u \perp \text{ span } (u_0, u_1, \ldots, u_{i-1})}} \frac{\int_\Omega |\nabla u|^2 dx}{\int_\Omega u^2 dx}. \tag{2.2}$$

(Here dx denotes standard Lebesgue measure in $I\!R^n$.) The key to the proof is to choose trial functions φ_i for each of the eigenfunctions u_i and insure that these are orthogonal to the preceding eigenfunctions $u_0, u_1, \ldots, u_{i-1}$. For the n trial functions $\varphi_1, \ldots, \varphi_n$ we simply choose

$$\varphi_j = x_j \quad \text{for } j = 1, \ldots, n, \tag{2.3}$$

but before we can use these we must make certain adjustments requiring two topological arguments. The first is to insure orthogonality of each of the φ_j's ($j = 1, \ldots, n$) to the eigenfunction u_0 (which is actually just the constant function $1/\sqrt{|\Omega|}$). This is done by using the Brouwer fixed point theorem precisely as done by Weinberger [44] (or see [10] or our papers [3], [4]). Simply by translating the origin appropriately we can guarantee $\varphi_j \perp u_0$ for $j = 1, \ldots, n$ (for the special case of $\varphi_j = x_j$ considered here this can be viewed simply as choosing to put the origin at the center of mass of Ω considered as a uniform mass distribution in $I\!R^n$). Next we argue, via the Borsuk–Ulam theorem (see [41, p. 266], [29, p. 170], or [31, p. 361]), that a suitable rotation of axes can be made so as to insure that

$$\int_\Omega \varphi_j u_i dx = \int_\Omega x_j u_i dx = 0 \tag{2.4}$$

for $j = 2, 3, \ldots, n$ and $i = 1, \ldots, j-1$. To see this, recall that the Borsuk–Ulam theorem says that a continuous mapping f from S^{n-1} (the unit sphere in $I\!R^n$) to $I\!R^{n-1}$, which is antipode preserving, i.e., $f(-\sigma) = -f(\sigma)$ for all $\sigma \in S^{n-1}$ must vanish somewhere (for $n \geq 2$). We apply this theorem first to the mapping $f_n : S^{n-1} \longrightarrow I\!R^{n-1}$ defined componentwise by

$$f_{n,k} : \sigma \longrightarrow \int_\Omega x_n u_k dx \quad \text{for } k = 1, \ldots, n-1, \tag{2.5}$$

where σ denotes the direction of the x_n axis (which we regard as free to vary over S^{n-1} by rotation; note that we leave the directions of the other axes ambiguous but that this has no effect on the values of the integrals in (2.5)). Since f_n is antipode preserving the Borsuk–Ulam theorem now tells us that there is a direction $\sigma_n \in S^{n-1}$ for which $f_n(\sigma_n) = 0$. We take this direction (its negative $-\sigma_n$ would serve equally well) as the direction of our x_n axis. This process is then repeated with functions $f_{n-1}, f_{n-2}, \ldots, f_2$ in sequence where the function $f_\ell : S^{\ell-1} \longrightarrow I\!R^{\ell-1}$ (for $\ell = 2, 3, \ldots,$ or $n-1$) is defined componentwise by

$$f_{\ell,k} : \sigma \longrightarrow \int_\Omega x_\ell u_k dx \quad \text{for } k = 1, 2, \ldots, \ell-1, \tag{2.6}$$

and σ is to be regarded as a variable for the direction of the x_ℓ axis and is free to range over all unit vectors perpendicular to the ones already fixed. That is, when considering f_ℓ we should consider its argument σ to vary over the sphere $S^{\ell-1}$ of values perpendicular to the directions $\sigma_{\ell+1}, \dots, \sigma_n$ of the axes $x_{\ell+1}, \dots, x_n$ which have been previously fixed in this inductive process. Once we choose σ_2, the direction of the x_2 axis, σ_1 is effectively fixed, and finally, we have reached the choice of axes that yields the orthogonality conditions (2.4). Note that the rotations we performed to get (2.4) cannot disrupt the orthogonality of each of the trial functions φ_j to the eigenfunction u_0 and thus, by (2.2), we arrive at

$$\mu_i \leq \frac{\int_\Omega |\nabla \varphi_i|^2 dx}{\int_\Omega \varphi_i^2 dx} = \frac{\int_\Omega dx}{\int_\Omega x_i^2 dx} = \frac{|\Omega|}{\int_\Omega x_i^2 dx} \tag{2.7}$$

for $i = 1, 2, \dots, n$. By inverting and summing on i we obtain

$$\sum_{i=1}^n \frac{1}{\mu_i} \geq \frac{\int_\Omega r^2 dx}{|\Omega|}, \tag{2.8}$$

where $r = |x|$ is the usual radial variable in spherical coordinates. Finally, it is easy to see by a simple rearrangement argument (see [37] or [44] for similar arguments) that

$$\int_\Omega r^2 dx \geq \int_{\Omega^*} r^2 dx, \tag{2.9}$$

where Ω^* denotes the n-dimensional ball of volume $|\Omega|$ having its center at the origin (Ω^* is called the spherical rearrangement of Ω). Since Ω^* has radius r_* where $|\Omega| = C_n r_*^n$ it is now easily computed that

$$\int_{\Omega^*} r^2 dx = nC_n \int_0^{r_*} r^{n+2-1} dr = \frac{n}{n+2} |\Omega| \left(\frac{|\Omega|}{C_n} \right)^{2/n}, \tag{2.10}$$

and putting this together with (2.8) and (2.9) immediately yields (2.1). □

Remark. Our application of the Borsuk–Ulam theorem above also occurred in our proof of (1.10) (see [5]) and in an argument in [6]. Both of these results concerned Dirichlet eigenvalues. Also, rearrangement arguments paralleling those given above occur in [37], [44] as already mentioned and also in work of Chiti [14], [15] and ourselves [3], [4], [6], where these references are roughly listed in order of complication (the first two deal with Neumann eigenvalues while the last five deal with Dirichlet eigenvalues; [44] and [3], [4], [6] give optimal bounds and hence involve Bessel functions as trial functions while the others use simpler trial functions). It might also be noted that in the two-dimensional case an argument leading to (2.4) could be made directly, without recourse to the Borsuk–Ulam theorem.

3. Asymptotics for large dimension.

In this section we present the following theorem, which shows that for a ball in n dimensions the ratio of the two sides of (2.1) approaches 1 as $n \to \infty$. More precisely, we have the following.

THEOREM 3.1. *For an n-dimensional ball,*

$$\frac{\left(\frac{1}{\mu_1} + \cdots + \frac{1}{\mu_n} \right)}{\left[n \left(|\Omega| / C_n \right)^{2/n} / (n+2) \right]} = 1 + O\left(\frac{1}{n} \right). \tag{3.1}$$

This theorem is a direct consequence of bounds for the zeros $p_{\nu,1}$ of $[x^{1-\nu}J_\nu(x)]'$ proved recently by Lorch and Szego [27]. The left-hand side of (3.1) reduces simply to

$$\frac{n+2}{p^2_{n/2,1}} \tag{3.2}$$

and (3.1) now follows immediately from the inequalities

$$2\nu + \frac{4}{\nu+3} < p^2_{\nu,1} < 2\nu + 2 \quad \text{for } \nu > -1 \tag{3.3}$$

of Lorch and Szego. We note that the upper bound in (3.3) for $\nu = n/2$ and $n = 1, 2, 3, \ldots$ is already a consequence of our inequality (2.1).

4. Bounds for domains with rotational symmetry. We begin by considering simply connected domains in two dimensions since slightly more detailed results can be obtained for them. We then turn back to general domains in n dimensions and obtain results for $\mu_1 + \cdots + \mu_n$ that are analagous to those we found earlier for $(\lambda_2 + \cdots + \lambda_{n+1})/\lambda_1$ (see §3 of [5]), where the λ_i's here represent the Dirichlet eigenvalues of $-\Delta$ on Ω. In addition, in two dimensions we shall obtain the bound $A\mu_2 \leq \pi(j'_{1,1})^2$ for any domain $\Omega \subset I\!R^2$ with 4-fold rotational symmetry and smooth boundary.

If a simply connected domain in two dimensions has a k-fold rotational symmetry about a point (which we take as our origin) with $k \geq 3$, then it follows relatively easily that $\mu_2 = \mu_1$ so that Weinberger's bound may be applied to produce various strengthenings of the previous results in this paper.

LEMMA 4.1. *If $\Omega \subset I\!R^2$ is a simply connected domain with k-fold symmetry where $k \geq 3$, then $\mu_2 = \mu_1$.*

Proof. Assume that Ω has k-fold rotational symmetry where $k \geq 3$, and let R_k denote the operation of rotating by $2\pi/k$ radians in the counterclockwise sense. Let u_1 be a real eigenfunction for μ_1. Then $u_1(R_kx)$ is also an eigenfunction for μ_1 and one of two conditions must hold: either (1) $u_1(R_kx)$ is a constant multiple of $u_1(x)$ or (2) $u_1(R_kx)$ and $u_1(x)$ are linearly independent eigenfunctions for μ_1. We now show that case (1) cannot occur, and hence that μ_1 is an eigenvalue of multiplicity at least two. Since Ω is simply connected, a well-known observation of Pleijel [36] (or see, for example, [10, p. 128], [19, p. 510], [21, p. 46], [33, p. 466]) shows that u_1 must have a crossing nodal line that is non–self-intersecting and meets the boundary $\partial\Omega$ at exactly two points, x_1 and x_2. This follows simply from the fact that if u_1 had a closed nodal line (the only other possibility), then we would have a domain Ω' (the nodal domain inside the closed nodal line) strictly smaller than Ω but with a smaller first Dirichlet eigenvalue (recall Pólya's result $\mu_1 < \lambda_1$), in contradiction to the well-known monotonicity of eigenvalues property (see [10], for example). In case (1) it is clear that the k points $R^i_kx_1$, for $i = 0, 1, \ldots, k-1$, must all be points where the nodal line of u_1 meets $\partial\Omega$ (as must the points $R^i_kx_2$ for $i = 0, 1, \ldots, k-1$). But this is impossible since the points $R^i_kx_1$ for $i = 0, 1, \ldots, k-1$ are all distinct and there are k of them with $k \geq 3$, in contradiction to the fact that the nodal line meets the boundary only at the two points x_1 and x_2. □

With Lemma 4.1 in hand the following inequalities are immediate (for a simply connected domain with k-fold rotational symmetry where $k \geq 3$ and with a smooth boundary):

$$\mu_2 \leq \frac{\pi(j'_{1,1})^2}{A}, \tag{4.1}$$

$$A(\mu_1 + \mu_2) \le 2\pi(j'_{1,1})^2, \tag{4.2}$$

and

$$\frac{1}{\mu_1} + \frac{1}{\mu_2} \ge \frac{2A}{\pi(j'_{1,1})^2}. \tag{4.3}$$

These are, of course, optimal results but (4.1) and (4.2), at least, could not be expected to continue to apply in the absence of k-fold symmetry (see §5). Also, (4.3) is just a special case of Szegö's result (1.16).

Under the assumption that Ω is 4-fold rotationally symmetric, but without assuming that Ω is simply connected or using facts about nodal lines, we can still prove (4.1), (4.2), and (4.3) in two dimensions. The only result above that cannot be proved in this context is the equality $\mu_2 = \mu_1$ of Lemma 4.1. We can also prove analogues of (4.2) and (4.3) in n dimensions. We begin with these since this is the most general part of the argument.

In n dimensions we consider domains Ω situated with respect to Cartesian coordinate axes such that with respect to rotation in each coordinate plane (i.e., $\binom{n}{2}$ planes) Ω has rotational symmetry of order 4. We shall follow closely Weinberger's proof while incorporating our 4-fold symmetry hypotheses.

THEOREM 4.2. *Let Ω be a bounded domain in $I\!R^n$ with smooth boundary and suppose that Ω is invariant with respect to* 90° *rotations in the coordinate planes spanned by each pair of* (*Cartesian*) *coordinate axes. Then*

$$\sum_{i=1}^{n} \mu_i \le n\left(\frac{C_n}{|\Omega|}\right)^{2/n} p^2_{n/2,1} \tag{4.4}$$

and

$$\sum_{i=1}^{n} \frac{1}{\mu_i} \ge \frac{n\left(|\Omega|/C_n\right)^{2/n}}{p^2_{n/2,1}}. \tag{4.5}$$

Proof. One works with trial functions $\varphi_j = g(r)x_j/r$ for $j = 1, \dots, n$ as introduced by Weinberger [44]. By the symmetry of Ω, orthogonality to u_0 is assured. But now one would like to use the extended Rayleigh–Ritz inequality which, in the special case in which we need it, asserts that

$$\sum_{j=1}^{n} \mu_j \le \sum_{j=1}^{n} \frac{\int_\Omega |\nabla \varphi_j|^2 dx}{\int_\Omega \varphi_j^2 dx} \tag{4.6}$$

if the φ_j's are pairwise orthogonal (and each is orthogonal to u_0). To see that these orthogonality conditions hold, the integrals $\int_\Omega \varphi_i\varphi_j dx = \int_\Omega [g(r)^2/r^2] x_i x_j dx$ for $i \ne j$ are investigated. Under a change of variables by a 90° rotation in the x_ix_j-coordinate plane this integral is seen to change to its negative and hence it must vanish. Thus

$$\int_\Omega \varphi_i\varphi_j dx = 0 \quad \text{for } i \ne j, \quad 1 \le i, \quad j \le n, \tag{4.7}$$

and use of (4.6) is justified. Before we can proceed, though, we also need to show that the integrals $\int_\Omega \varphi_j^2 dx$ are the same for all $j = 1, 2, \dots, n$. This again follows easily by a 90° rotational change of variables applied to the integral $\int_\Omega [\varphi_i^2 - \varphi_j^2]dx$. Thus the denominators in the right-hand side of (4.6) are all equal (as are the numerators by a similar argument). Hence we can sum to arrive at

$$\sum_{j=1}^{n} \mu_j \leq \frac{\sum_{j=1}^{n} \int_\Omega |\nabla \varphi_j|^2 dx}{(1/n) \sum_{j=1}^{n} \int_\Omega \varphi_j^2 dx} = \frac{n \int_\Omega [g'(r)^2 + (n-1)g(r)^2/r^2]\, dx}{\int_\Omega g(r)^2 dx}. \tag{4.8}$$

If we now make Weinberger's choice of function $g(r)$ (this would be $G(r)$ in his notation) we obtain (4.4), as desired. Inequality (4.5) follows from the Cauchy–Schwarz inequality applied as follows:

$$n^2 = \left(\sum_{j=1}^{n} \frac{1}{\sqrt{\mu_j}} \sqrt{\mu_j} \right)^2 \leq \left(\sum_{j=1}^{n} \frac{1}{\mu_j} \right) \left(\sum_{j=1}^{n} \mu_j \right). \quad \square$$

In two dimensions we can go beyond Theorem 4.2 to the following.

THEOREM 4.3. *Let $\Omega \subset I\!R^2$ be a bounded domain having 4-fold rotational symmetry. Then*

$$\mu_j \leq \frac{\pi \left(j'_{1,1}\right)^2}{A} \quad \textit{for } j = 1, 2, \tag{4.9}$$

and hence also $A(\mu_1 + \mu_2) \leq 2\pi \left(j'_{1,1}\right)^2$.

Proof. We develop separate inequalities for μ_1 and μ_2. To do this we use the variational characterizations of μ_1 and μ_2. By the 4-fold symmetry of Ω we can insist that u_1 be either even or odd under rotation by 180°. We use this together with rotational freedom to argue that we can make a choice of axes so that $\int_\Omega \varphi_2 u_1 \, dx = 0$. If u_1 were even, then by employing a 180°-rotational change of variables in the integral $\int_\Omega \varphi_2 u_1 \, dx$ we find that it equals its own negative and is hence zero. If u_1 were odd, then a 180° rotation of the coordinate system reverses the sign of $\int_\Omega \varphi_2 u_1 \, dx$, implying that for some choice of rotated coordinates $\int_\Omega \varphi_2 u_1 \, dx$ vanishes. We then obtain

$$\mu_j \leq \frac{\int_\Omega |\nabla \varphi_j|^2 \, dx}{\int_\Omega \varphi_j^2 \, dx} \quad \text{for } j = 1, 2, \tag{4.10}$$

which in turn yields

$$\mu_j \leq \frac{\int_\Omega [g'(r)^2 + (n-1)g(r)^2/r^2] \, dx}{\int_\Omega g(r)^2 \, dx}, \tag{4.11}$$

since as before neither $\int_\Omega \varphi_j^2 \, dx$ nor $\int_\Omega |\nabla \varphi_j|^2 \, dx$ varies with j. Weinberger's choice of function $g(r)$ then yields the inequalities (4.9) of the theorem. □

If u_1 happens to be even, then our proof above actually yields more.

THEOREM 4.4. *With hypotheses as in Theorem 4.3 suppose, in addition, that u_1 is even, i.e., that $u_1(-x) = u_1(x)$ for all $x \in \Omega$. Then*

$$\mu_j \leq \frac{\pi (j'_{1,1})^2}{A} \quad \textit{for } j = 2, 3, \tag{4.12}$$

and hence also $A(\mu_2+\mu_3) \le 2\pi(j_{1,1}')^2$.

For the proof we simply use the argument above which led to (4.9) but noting that $\int_\Omega \varphi_i u_1\,dx = 0$ for $i = 1,2$ is automatic (since u_1 is even) so that we can, by making a suitable rotation, view φ_1 and φ_2 as trial functions for u_2 and u_3, respectively.

Theorem 4.4 has potential relevance for nonsimply connected domains. One can imagine a domain in the shape of a gasket having 4 holes and which is 4-fold rotationally invariant. If the holes are elongated and come close to disconnecting the domain into a disk and an annulus (of appropriate radii), then it seems reasonable to expect that u_1 will be 4-fold rotationally invariant with its nodal set consisting of 4 short curves connecting the holes.

5. Examples in two dimensions. For rectangles (always assuming an $a \times b$ rectangle with $a \ge b > 0$) we obtain

$$\mu_1 = \frac{\pi^2}{a^2}, \tag{5.1}$$

$$\mu_2 = \begin{cases} \dfrac{\pi^2}{b^2} & \text{if } \dfrac{1}{2} \le \dfrac{b}{a} \le 1, \\ \dfrac{4\pi^2}{a^2} & \text{if } 0 < \dfrac{b}{a} \le \dfrac{1}{2}. \end{cases} \tag{5.2}$$

Thus

$$A\mu_1 = \frac{\pi^2 b}{a}, \tag{5.3}$$

$$A\mu_2 = \begin{cases} \dfrac{\pi^2 a}{b} & \text{if } \dfrac{1}{2} \le \dfrac{b}{a} \le 1, \\ \dfrac{4\pi^2 b}{a} & \text{if } 0 < \dfrac{b}{a} \le \dfrac{1}{2}, \end{cases} \tag{5.4}$$

$$A(\mu_1+\mu_2) = \begin{cases} \pi^2\left(\dfrac{b}{a}+\dfrac{a}{b}\right) & \text{if } \dfrac{1}{2} \le \dfrac{b}{a} \le 1, \\ 5\pi^2 b/a & \text{if } 0 < \dfrac{b}{a} \le \dfrac{1}{2}, \end{cases} \tag{5.5}$$

$$\frac{1}{A}\left(\frac{1}{\mu_1}+\frac{1}{\mu_2}\right) = \begin{cases} \dfrac{\left(\dfrac{b}{a}+\dfrac{a}{b}\right)}{\pi^2} & \text{if } \dfrac{1}{2} \le \dfrac{b}{a} \le 1, \\ \dfrac{5a}{4\pi^2 b} & \text{if } 0 < \dfrac{b}{a} \le \dfrac{1}{2}, \end{cases} \tag{5.6}$$

and

$$\frac{\mu_2}{\mu_1} = \begin{cases} \dfrac{a^2}{b^2} & \text{if } \dfrac{1}{2} \le \dfrac{b}{a} \le 1, \\ 4 & \text{if } 0 < \dfrac{b}{a} \le \dfrac{1}{2}. \end{cases} \tag{5.7}$$

From these it follows that among rectangles $A\mu_2$ and $A(\mu_1+\mu_2)$ take their maxima for the 2:1 rectangle. Also $(1/\mu_1 + 1/\mu_2)/A$ takes its minimum among rectangles at the square.

Note that the 2:1 rectangle shows that the disk does not maximize $A(\mu_1+\mu_2)$ (for the disk $A(\mu_1+\mu_2) \approx 21.2996$). Of course, Weinberger's result guarantees that $A\mu_1$ is maximized at the disk. It seems likely, too, that the disk minimizes $(1/\mu_1 + 1/\mu_2)/A$. We also hazard the guess that $\mu_2/\mu_1 \leq 4$ for all *convex* domains. Finally, we note that our inequality (1.17) provides the lower bound $1/2\pi \approx .15915$ to $(1/\mu_1 + 1/\mu_2)/A$, which would be only about 15 percent low if the optimal bound were its value at the disk($\approx .18780$).

Acknowledgment. We would like to thank Hans Weinberger for several useful comments and, in particular, for a more general formulation of Lemma 4.1. We are also grateful to Lee Lorch and Peter Szego for sharing the results in [27] with us prior to publication.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, National Bureau of Standards Applied Mathematics Series, 55 (1964), U.S. Government Printing Office, Washington, DC.

[2] V. I. Arnol'd, M. I. Vishik, Yu. S. Il'yashenko, A. S. Kalashnikov, V. A. Kondrat'ev, S. N. Kruzhkov, E. M. Landis, V. M. Millionshchikov, O. A. Oleinik, A. F. Filippov, and M. A. Shubin, *Some unsolved problems in the theory of differential equations and mathematical physics*, Russian Math. Surveys, 44 (1989), pp. 157–171. (Translation of Uspekhi Mat. Nauk, 44 (1989), pp. 191–202.)

[3] M. S. Ashbaugh and R. D. Benguria, *Proof of the Payne–Pólya–Weinberger conjecture*, Bull. Amer. Math. Soc., 25 (1991), pp. 19–29.

[4] ———, *A sharp bound for the ratio of the first two eigenvalues of Dirichlet Laplacians and extensions*, Ann. Math., 135 (1992), pp. 601–628.

[5] ———, *More bounds on eigenvalue ratios for Dirichlet Laplacians in n dimensions*, 1991, preprint.

[6] ———, *Isoperimetric bound for λ_3/λ_2 for the membrane problem*, Duke Math. J., 63 (1991), pp. 333–341.

[7] ———, *Isoperimetric bounds for higher eigenvalue ratios for the n-dimensional fixed membrane problem*, Proc. Roy. Soc. Edinburgh, to appear.

[8] R. Askey, ed., *Gabor Szegö: Collected Papers, Vol.* 3: 1945–1972, Birkhäuser, Boston, MA, 1982.

[9] P. Aviles, *Symmetry theorems related to Pompeiu's problem*, Amer. J. Math., 108 (1986), pp. 1023–1036.

[10] C. Bandle, *Isoperimetric Inequalities and Applications*, Pitman Monographs and Studies in Math., 7, Pitman, Boston, 1980.

[11] J. J. A. M. Brands, *Bounds for the ratios of the first three membrane eigenvalues*, Arch. Rational Mech. Anal., 16 (1964), pp. 265–268.

[12] I. Chavel, *Lowest eigenvalue inequalities*, in Geometry of the Laplace Operator, Proc. Sympos. Pure Math., Vol. 36, S. S. Chern and A. Weinstein, eds., American Mathematical Society, Providence, RI, 1980, pp. 79–89.

[13] ———, *Eigenvalues in Riemannian Geometry*, Academic, New York, 1984.

[14] G. Chiti, *Inequalities for the first three membrane eigenvalues*, Boll. Un. Mat. Ital. (Series 5), 18-A (1981), pp. 144–148.

[15] ———, *A bound for the ratio of the first two eigenvalues of a membrane*, SIAM J. Math. Anal., 14 (1983), pp. 1163–1167.

[16] Y. Colin de Verdière, *Construction de Laplaciens dont une partie finie du spectre est donnée*, Ann. Scient. École Norm. Sup. (Series 4), 20 (1987), pp. 599–615.

[17] G. Faber, *Beweis, dass unter allen homogenen Membranen von gleicher Fläche und gleicher Spannung die kreisförmige den tiefsten Grundton gibt*, Sitzungberichte der mathematisch-physikalischen Klasse der Bayerischen Akademie der Wissenschaften zu München Jahrgang, 1923, pp. 169–172.

[18] L. FRIEDLANDER, *Some inequalities between Dirichlet and Neumann eigenvalues*, Arch. Rational Mech. Anal., 116 (1991), pp. 153–160.

[19] J. HERSCH AND G.-C. ROTA, EDS., *George Pólya: Collected Papers, Vol.* III: *Analysis*, MIT Press, Cambridge, MA, 1984.

[20] G. N. HILE AND M. H. PROTTER, *Inequalities for eigenvalues of the Laplacian*, Indiana Univ. Math. J., 29 (1980), pp. 523–538.

[21] B. KAWOHL, *Rearrangements and Convexity of Level Sets in PDE*, Lecture Notes in Math., 1150, Springer-Verlag, Berlin, 1985.

[22] E. T. KORNHAUSER AND I. STAKGOLD, *A variational theorem for $\nabla^2 u + \lambda u = 0$ and its applications*, J. Math. Phys., 31 (1952), pp. 45–54.

[23] E. KRAHN, *Über eine von Rayleigh formulierte Minimaleigenschaft des Kreises*, Math. Ann., 94 (1925), pp. 97–100.

[24] ———, *Über Minimaleigenschaften der Kugel in drei und mehr Dimensionen*, Acta Comm. Univ. Tartu (Dorpat) A, 9 (1926), pp. 1–44.

[25] H. A. LEVINE, *Some remarks on inequalities between Dirichlet and Neumann eigenvalues*, in Maximum Principles and Eigenvalue Problems in Partial Differential Equations, P. W. Schaefer, ed., Pitman Res. Notes in Math. 175, Longman Scientific and Technical, Harlow, Essex, United Kingdom, 1988, pp. 121–133.

[26] H. A. LEVINE AND H. F. WEINBERGER, *Inequalities between Dirichlet and Neumann eigenvalues*, Arch. Rational Mech. Anal., 94 (1986), pp. 193–208.

[27] L. LORCH AND P. SZEGO, *Bounds and monotonicities for the zeros of derivatives of ultraspherical Bessel functions*, 1992, preprint.

[28] P. MARCELLINI, *Bounds for the third membrane eigenvalue*, J. Differential Equations, 37 (1980), pp. 438–443.

[29] W. S. MASSEY, *Algebraic Topology: An Introduction*, Harcourt, Brace and World, New York, 1967.

[30] R. MAZZEO, *Remarks on a paper of Friedlander concerning inequalities between Neumann and Dirichlet eigenvalues*, Internat. Math. Res. Notices, 4 (1991), pp. 41–48.

[31] J. R. MUNKRES, *Topology, A First Course*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

[32] L. E. PAYNE, *Inequalities for eigenvalues of membranes and plates*, J. Rational Mech. Anal., 4 (1955), pp. 517–529.

[33] ———, *Isoperimetric inequalities and their applications*, SIAM Review, 9 (1967), pp. 453–488.

[34] L. E. PAYNE, G. PÓLYA, AND H. F. WEINBERGER, *Sur le quotient de deux fréquences propres consécutives*, Comptes Rendus Acad. Sci. Paris, 241 (1955), pp. 917–919 (reprinted as pp. 410–412 of [19] with comments by J. Hersch on p. 518).

[35] ———, *On the ratio of consecutive eigenvalues*, J. Math. Phys., 35 (1956), pp. 289–298 (reprinted as pp. 420–429 of [19] with comments by J. Hersch on pp. 521–522).

[36] A. PLEIJEL, *Remarks on Courant's nodal line theorem*, Comm. Pure Appl. Math., 9 (1956), pp. 543–550.

[37] G. PÓLYA, *Remarks on the foregoing paper*, J. Math. Phys., 31 (1952), pp. 55–57 (reprinted as pp. 270–272 of [19] with comments by J. Hersch on pp. 509–510).

[38] ———, *On the characteristic frequencies of a symmetric membrane*, Math. Z., 63 (1955), pp. 331–337 (reprinted as pp. 413–419 of [19] with comments by J. Hersch on pp. 519–521).

[39] M. H. PROTTER, *Can one hear the shape of a drum? revisited*, SIAM Review, 29 (1987), pp. 185–197.

[40] ———, *Universal inequalities for eigenvalues*, in Maximum Principles and Eigenvalue Problems in Partial Differential Equations, P. W. Schaefer, ed., Pitman Res. Notes in Math. 175, Longman Scientific and Technical, Harlow, Essex, United Kingdom, 1988, pp. 111–120.

[41] E. H. SPANIER, *Algebraic Topology*, McGraw-Hill, New York, 1966.

[42] G. SZEGÖ, *Inequalities for certain eigenvalues of a membrane of given area*, J. Rational Mech. Anal., 3 (1954), pp. 343–356 (reprinted as pp. 373–386 of [8] with comments by R. Askey on p. 387).

[43] C. J. THOMPSON, *On the ratio of consecutive eigenvalues in N-dimensions*, Stud. Appl. Math., 48 (1969), pp. 281–283.

[44] H. F. WEINBERGER, *An isoperimetric inequality for the n-dimensional free membrane problem*, J. Rational Mech. Anal., 5 (1956), pp. 633–636.

SIAM J. MATH. ANAL.
Vol. 24, No. 3, pp. 571–582, May 1993

AN EXISTENCE THEOREM FOR A FREE BOUNDARY PROBLEM OF HYPERSONIC FLOW THEORY*

VINCENT GIOVANGIGLI†

Abstract. The author considers a free boundary problem arising in hypersonic flow theory. The model describes an axisymmetric thin viscous shock layer in the vicinity of the stagnation point of a blunt body. The governing equations on the symmetry line reduce to a two-point boundary value problem with four unknowns and a free boundary. The free boundary problem is reduced to a nonlinear eigenvalue problem through a change of variable. Existence of a solution is achieved by deriving a priori estimates and by using the Leray–Schauder topological degree theory.

Key words. free boundary, viscous shock layer, existence

AMS subject classifications. 65L10, 76N99

1. Introduction.

An interesting free boundary problem arising in hypersonic flow theory is that of thin viscous shock layers. In these models, the governing equations are of compressible boundary layer type with a normal pressure gradient and are solved between the body surface and the shock. Usual boundary conditions are imposed at the body wall, whereas the Rankine–Hugoniot relations are written at the shock which is the free boundary. These models have been found successful is predicting hypersonic flows past blunt bodies; for more details we refer to Blottner [2], Bush [3], Davis [5], Ho and Probstein [7], Laboudigue, Giovangigli, and Candel [8], and to the references therein.

In this paper, we investigate an axisymmetric thin viscous shock layer for a perfect gas in the vicinity of the stagnation point of a blunt body. Only the governing equations on the symmetry line are considered. The corresponding solutions provide the initial conditions for a finite difference method proceeding in the downstream direction [2], [5], [8] as it has been shown rigorously by Oleinik [10] in the incompressible flat boundary layer case. Another point of view is that these solutions are similar approximated solutions that are valid in the neighborhood of the stagnation streamline as are the solutions of the Falkner–Skan equations for incompressible boundary layers.

The thin viscous shock layer governing equations along the symmetry line reduce to a two-point boundary value problem with four unknowns and with a free boundary. The four reduced solution components are the normal and tangential velocities, the pressure, and the temperature. The tangential velocity and the temperature are governed by second-order equations and are specified at the boundaries. On the other hand, the normal velocity and the pressure are governed by first-order equations, but there are two boundary conditions for the normal velocity and one for the pressure. The extra boundary condition for the normal velocity is thus used to determine the free boundary. By introducing a reduced normal coordinate, the free boundary problem is first transformed into a nonlinear eigenvalue problem. Existence of a solution is then achieved by deriving a priori estimates and by using the Leray–Schauder topological degree theory. The main difficulty consists in deriving a priori estimates of the solution components. In particular, obtaining strictly positive upper and lower bounds for

*Received by the editors November 13, 1991; accepted for publication (in revised form) August 13, 1992.

†Centre de Mathématiques Appliquées and Centre National de la Recherche Scientifique, Ecole Polytechnique 91128 Palaiseau cedex, France.

the free boundary is essential and rely on the presence of a negative pressure gradient term in the tangential momentum equation. The method of proof is similar to the one used by Berestycki, Nicolaenko, and Scheurer for flame problems [1], [6]. This proof differs from those of Weyl [11] and McLeod and Serrin [9] in the case of similar flat boundary layers which were based on shooting techniques. It is not known whether such a solution is unique as in the compressible boundary layer case [9] at variance with the incompressible boundary layer case [4]. Finally, to the author's knowledge, the results that are presented in this paper are new.

The thin viscous shock layer equations are presented in §2 for self-completeness. A priori estimates and existence of a solution are obtained in §3.

2. Setting of the problem. In this section, for self-completeness, we derive the first-order axisymmetric thin viscous shock layer equations on the symmetry line. For more details on the derivation of the model, we refer to Blottner [2], Bush [3], Davis [5], Ho and Probstein [7], Laboudigue, Giovangigli, and Candel [8], and to the references therein.

2.1. Viscous shock layers. The axisymmetric viscous shock layer equations can be derived by writing the Navier–Stokes equations in a local coordinate system (s,n), where s is measured along the body surface and n normal to the body surface. The shock Reynolds number Re_σ is assumed to be large $Re_\sigma = \rho_\infty V_\infty d/\mu_\sigma \gg 1$, where ρ_∞ denotes the density in front of the shock, V_∞ the uniform velocity in front of the shock, d the radius of curvature at the nose of the blunt body, and μ_σ the viscosity behind the shock, and the order of magnitude of the terms in the equations is determined with respect to the inverse square root of Re_σ. Only first-order terms are kept from both a viscous and an inviscid view point, so that the governing equations are uniformly valid between the shock and the body surface. The total mass conservation equation then takes the form

$$\frac{\partial}{\partial s}(\mathcal{R}\rho U) + \frac{\partial}{\partial n}(\mathcal{H}\mathcal{R}\rho V) = 0, \tag{2.1}$$

where ρ is the density, $\mathcal{H} = 1 + \kappa n$ a curvature term, κ the positive curvature of the body surface, $\mathcal{R}$ the radius from the axis of symmetry, (U,V) the mass averaged flow velocity components in the local coordinate system, and the tangential momentum conservation equation is

$$\frac{\rho U}{\mathcal{H}}\frac{\partial U}{\partial s} + \rho V\frac{\partial U}{\partial n} + \frac{\rho\kappa UV}{\mathcal{H}} + \frac{1}{\mathcal{H}}\frac{\partial P}{\partial s} - \frac{1}{\mathcal{H}\mathcal{R}}\frac{\partial}{\partial n}\Big(\mathcal{H}\mathcal{R}\mu\frac{\partial U}{\partial n}\Big) = 0, \tag{2.2}$$

where P is the pressure and μ the viscosity. The normal momentum conservation can also be written

$$\frac{\rho U}{\mathcal{H}}\frac{\partial V}{\partial s} + \rho V\frac{\partial V}{\partial n} - \frac{\rho\kappa U^2}{\mathcal{H}} + \frac{\partial P}{\partial n} = 0, \tag{2.3}$$

and the energy conservation equation is

$$\rho c_p\frac{U}{\mathcal{H}}\frac{\partial T}{\partial s} + \rho c_p V\frac{\partial T}{\partial n} - \frac{U}{\mathcal{H}}\frac{\partial P}{\partial s} - V\frac{\partial P}{\partial n} - \mu\Big(\frac{\partial U}{\partial n}\Big)^2 - \frac{1}{\mathcal{H}\mathcal{R}}\frac{\partial}{\partial n}\Big(\mathcal{H}\mathcal{R}k\frac{\partial T}{\partial n}\Big) = 0, \tag{2.4}$$

where T is the temperature, c_p the constant pressure specific heat—assumed to be a constant—and k the thermal conductivity. Furthermore, the state law reads

$$\rho = \frac{\gamma}{\gamma - 1}\frac{P}{c_p T}, \tag{2.5}$$

where γ denotes the ratio of specific heats $\gamma = c_p/c_v$.

The boundary conditions at the shock are provided by the Rankine–Hugoniot relations in the high Mach number limit $M_\infty = V_\infty^2/(\gamma P_\infty) \gg 1$, where P_∞ denotes the uniform pressure in front of the shock. Denoting by $n = \sigma(s)$ the shock location, we have

$$V\big(s,\sigma(s)\big) = V_\infty \cos(\phi+\theta)\sin\theta - \epsilon V_\infty \sin(\phi+\theta)\cos\theta, \tag{2.6}$$

$$U\big(s,\sigma(s)\big) = V_\infty \cos(\phi+\theta)\cos\theta + \epsilon V_\infty \sin(\phi+\theta)\sin\theta, \tag{2.7}$$

$$P\big(s,\sigma(s)\big) = \frac{2\rho_\infty V_\infty^2}{\gamma+1}\sin^2(\phi+\theta), \qquad T\big(s,\sigma(s)\big) = \frac{2\gamma V_\infty^2}{(\gamma+1)^2 c_p}\sin^2(\phi+\theta), \tag{2.8}$$

where $\epsilon = (\gamma-1)/(\gamma+1)$ and where ρ_∞ denotes the density of air in front of the shock. Remark also that from (2.5) and (2.8) we have $\rho\big(s,\sigma(s)\big) = \rho_\infty/\epsilon$. In these relations ϕ denotes the angle between the axis of symmetry and the tangent to the body surface and θ the angle between the tangent to the body surface and the tangent to the shock, i.e., $\tan\theta = (d\sigma/ds)/(1+\kappa\sigma)$. The radius $\mathcal{R}$ is also given by $\mathcal{R} = r + n\cos\phi$, where r denotes the distance from the axis of symmetry to the body surface and the positive curvature κ, the radius r and the angle ϕ are given function of s such that $-\kappa = d\phi/ds$ and $\sin\phi = dr/ds$. Futhermore, with $s = 0$ corresponding to the symmetry axis, i.e., the stagnation streamline, we have $\kappa(s) = 1/d + O(s^2)$, $\phi(s) = \pi/2 - s/d + O(s^3)$, and $r(s) = s + O(s^3)$, where $d = 1/\kappa(0)$ is the radius of curvature at the nose of the blunt body. On the other hand, at the body surface, we have

$$V(s,0) = 0, \qquad U(s,0) = 0, \qquad T(s,0) = T_w, \tag{2.9}$$

where T_w denotes the body wall temperature.

2.2. Thin viscous shock layers. The thin shock layer approximation is now used in order to simplify the system (2.1)–(2.9). In the thin layer limit, we have $\epsilon = (\gamma-1)/(\gamma+1) \ll 1$, which yields the following orders of magnitude $s/d = O(1)$, $n/d = O(\epsilon)$, $\rho/\rho_\infty = O(1/\epsilon)$, $U/V_\infty = O(1)$, $V/V_\infty = O(\epsilon)$, $P/(\rho_\infty V_\infty^2) = O(1)$, and $c_p T/V_\infty^2 = O(1)$. Note that in the thin layer approximation we still have $\epsilon Re_\sigma \gg 1$ and $\epsilon M_\infty^2 \gg 1$, although $\epsilon \ll 1$ [3], [7]. Under the thin layer approximation, we thus have $\mathcal{H}(s,n) = 1 + \kappa(s)n = 1 + O(\epsilon)$ and $\mathcal{R}(s,n) = r(s) + n\cos\phi(s) = r(s)\big(1 + O(\epsilon)\big)$ so that up to first-order we may replace $\mathcal{H}$ by 1 and $\mathcal{R}$ by r in (2.1)–(2.9). An examination of the normal momentum equation also indicates that the terms $\rho U(\partial V/\partial s)$ and $\rho V(\partial V/\partial n)$ are of lower order and can thus be neglected. Similarly, the terms $U(\partial P/\partial s)$ and $V(\partial P/\partial n)$ can be eliminated from the energy conservation equation. The first-order axisymmetric thin viscous shock layer equations are thus found to be

$$\frac{1}{r}\frac{\partial(r\rho U)}{\partial s} + \frac{\partial(\rho V)}{\partial n} = 0, \tag{2.10}$$

$$\rho U\frac{\partial U}{\partial s} + \rho V\frac{\partial U}{\partial n} + \kappa\rho UV + \frac{\partial P}{\partial s} - \frac{\partial}{\partial n}\Big(\mu\frac{\partial U}{\partial n}\Big) = 0, \tag{2.11}$$

$$-\kappa\rho U^2+\frac{\partial P}{\partial n}=0, \tag{2.12}$$

$$\rho c_p U\frac{\partial T}{\partial s}+\rho c_p V\frac{\partial T}{\partial n}-\mu\left(\frac{\partial U}{\partial n}\right)^2-\frac{\partial}{\partial n}\left(k\frac{\partial T}{\partial n}\right)=0. \tag{2.13}$$

Note that the curvature term $\kappa\rho UV$ is usually neglected in the tangential momentum equation, but is kept here for consistency with the curvature term $\kappa\rho U^2$ in the normal momentum equation. The boundary conditions (2.6)–(2.8) can also be simplified since we now have $\sigma/d=O(\epsilon)$ so that $\cos\theta=1+O(\epsilon^2)$ and $\sin\theta=d\sigma/ds+O(\epsilon^2)=O(\epsilon)$, which yields

$$V\big(s,\sigma(s)\big)-U\big(s,\sigma(s)\big)\frac{d\sigma}{ds}(s)=-\epsilon V_\infty\sin\phi(s),\qquad U\big(s,\sigma(s)\big)=V_\infty\cos\phi(s), \tag{2.14}$$

$$P\big(s,\sigma(s)\big)=\frac{2\rho_\infty V_\infty^2}{(\gamma+1)}\sin^2\phi(s),\qquad T\big(s,\sigma(s)\big)=\frac{2\gamma V_\infty^2}{(\gamma+1)^2c_p}\sin^2\phi(s), \tag{2.15}$$

and the values of ρ, V, U, P, T at the shock will be denoted by ρ_σ, V_σ, U_σ, P_σ, and T_σ, respectively, in the following. On the other hand, the boundary conditions (2.9) at the body surface are unchanged. Finally note that $V_\sigma-U_\sigma(d\sigma/ds)$, T_σ, P_σ, and T_w are even functions of s as are expected to be V, T, P, and σ, whereas U_σ, r, and $\pi/2-\phi$ are odd functions of s as is expected to be U.

2.3. Transformed equations. We now recast the governing equations into a simpler form, and we simultaneously transform the free boundary problem into a nonlinear eigenvalue problem. Assuming a positive shock location σ, we may introduce the new independent variables (ξ,ζ)—similar to the Howarth–Dorodnitsyn variables [2]—defined by

$$\xi=\frac{S(s)}{|S(s)|}|S(s)|^{1/4},\qquad \zeta=\frac{1}{l}\frac{U_\sigma(s)r(s)}{\sqrt{2S(s)}}\int_0^n\rho(s,\nu)\,d\nu, \tag{2.16}$$

where S and l are given by

$$S=\int_0^s\rho_\sigma(t)\mu_\sigma(t)U_\sigma(t)r^2(t)dt,\qquad l=\frac{U_\sigma(s)r(s)}{\sqrt{2S(s)}}\int_0^{\sigma(s)}\rho(s,\nu)\,d\nu, \tag{2.17}$$

so that l is the new unknown associated with the shock location σ. We also introduce the new dependent variables

$$v=\frac{\xi l}{2U_\sigma}\frac{\partial s}{\partial\xi}\left(U\frac{\partial\zeta}{\partial s}+V\frac{\partial\zeta}{\partial n}\right),\qquad u=\frac{1}{U_\sigma}\frac{\partial s}{\partial\xi}\left(U\frac{\partial\xi}{\partial s}+V\frac{\partial\xi}{\partial n}\right)=\frac{U}{U_\sigma}, \tag{2.18}$$

$$p=\frac{P-P(0,\sigma(0))}{\rho_\sigma U_\sigma^2},\qquad \theta=\frac{T}{T_\sigma}, \tag{2.19}$$

and after lengthy calculations, the transformed governing equations are shown to be

$$\frac{\xi}{2}\frac{\partial u}{\partial\xi}+u\left(1+\frac{\xi}{2l}\frac{\partial l}{\partial\xi}\right)+\frac{1}{l}\frac{\partial v}{\partial\zeta}=0, \tag{2.20}$$

$$\frac{\xi u}{2}\frac{\partial u}{\partial \xi}+\alpha u^2+\frac{\mathcal{K}uv}{\rho}+\frac{v}{l}\frac{\partial u}{\partial \zeta}+\frac{2\alpha p\rho_\sigma}{\rho}+\frac{\xi\rho_\sigma}{2\rho}\frac{\partial p}{\partial \xi}-\frac{1}{l^2}\frac{\partial}{\partial \zeta}\left(\eta\frac{\partial u}{\partial \zeta}\right)=0, \tag{2.21}$$

$$-\frac{\mathcal{K}u^2}{\rho_\sigma}+\frac{1}{l}\frac{\partial p}{\partial \zeta}=0, \tag{2.22}$$

$$\frac{\xi u}{2}\frac{\partial \theta}{\partial \xi}+\beta u\theta+\frac{v}{l}\frac{\partial \theta}{\partial \zeta}-e\eta\left(\frac{\partial u}{\partial \zeta}\right)^2-\frac{1}{l^2}\frac{\partial}{\partial \zeta}\left(\lambda\frac{\partial \theta}{\partial \zeta}\right)=0, \tag{2.23}$$

where α, β, e, and $\mathcal{K}$ are functions of ξ coefficients given by

$$\alpha=\frac{\xi}{2U_\sigma}\frac{\partial U_\sigma}{\partial \xi},\qquad \beta=\frac{\xi}{2T_\sigma}\frac{\partial T_\sigma}{\partial \xi},\qquad e=\frac{U_\sigma^2}{c_pT_\sigma},\qquad \mathcal{K}=\frac{\kappa\sqrt{2S}}{U_\sigma r},$$

and where the reduced transport coefficients η and λ are given by

$$\mu=\frac{\rho\mu}{\rho_\sigma\mu_\sigma},\qquad \lambda=\frac{\rho k}{\rho_\sigma\mu_\sigma c_p}.$$

In particular, keeping in mind that $\xi=\delta s+O(s^3)$ with $\delta>0$, in the neighborhood of $\xi=0$, we have $\alpha=\alpha(0)+O(\xi^2)$, $\beta=O(\xi^2)$, $e=O(\xi^2)$, and $\mathcal{K}=\mathcal{K}(0)+O(\xi^2)$, with $\alpha(0)>0$ and $\mathcal{K}(0)>0$. The corresponding boundary conditions for $\zeta=1$ are now

$$v(\xi,1)=v_\sigma(\xi),\qquad u(\xi,1)=u_\sigma(\xi),\qquad p(\xi,1)=p_\sigma(\xi),\qquad \theta(\xi,1)=\theta_\sigma(\xi), \tag{2.24}$$

where $v_\sigma=\big(V_\sigma-U_\sigma(d\sigma/ds)\big)\sqrt{2S}/(U_\sigma\mu_\sigma r)$, $u_\sigma=1$, $p_\sigma=-2(\gamma-1)/(\gamma+1)^2$, and $\theta_\sigma=1$ are known functions of ξ, and at the body surface $\zeta=0$ we have

$$v(\xi,0)=0,\qquad u(\xi,0)=0,\qquad \theta(\xi,0)=\theta_w(\xi), \tag{2.25}$$

where $\theta_w=T_w/T_\sigma$ denotes the reduced body wall temperature. Remark that the coefficients α, β, e, $\mathcal{K}$, v_σ, u_σ, p_σ, θ_σ, and θ_w are now even functions of ξ as are expected to be the new unknowns v, u, p, θ, and l.

2.4. Equations along the symmetry axis and locally similar solutions. In this paper, we only investigate the solutions along the symmetry line $\xi=0$. These solutions provide the initial conditions for a finite difference method proceeding in the downstream direction [2], [5], [8] as it has been shown rigorously by Oleinik in the incompressible flat boundary layer case [10]. These solutions on the stagnation streamline $\xi=0$ are also locally similar approximated solutions that are valid in the neighborhood of the symmetry line as are the solutions of the Falkner–Skan equations for incompressible boundary layers. More specifically, replacing all the coefficients by their constant quadratic approximations in (2.20)–(2.25) yields that $\beta=0$ and $e=0$ and that α, $\mathcal{K}$, v_σ, u_σ, p_σ, θ_σ, and θ_w are constants. But then the resulting system of partial differential equations posseses similar solutions functions of ζ alone, i.e., such that $\partial/\partial\xi=0$, and these similar solutions are governed by the same equations as the one that are strictly valid along the stagnation streamline $\xi=0$ since only the grouping $\xi(\partial/\partial\xi)$ appears in the governing equations (2.20)–(2.25).

Since now we only consider the solutions along the stagnation streamline $\xi=0$, we formally drop the ξ dependence in order to avoid notational complexity. Therefore,

v, u, p, θ, and l now stand, respectively, for $v(0,\cdot)$, $u(0,\cdot)$, $p(0,\cdot)$, $\theta(0,\cdot)$, and $l(0)$ and α, v_σ, u_σ, p_σ, θ_σ, and θ_w stand, respectively, for $\alpha(0)$, $v_\sigma(0)$, $u_\sigma(0)$, $p_\sigma(0)$, $\theta_\sigma(0)$, and $\theta_w(0)$. Denoting by $'$ the derivation with respect to ζ, the equations on the symmetry line $\xi = 0$ are then

$$u + \frac{v'}{l} = 0, \tag{2.26}$$

$$\alpha u^2 + cuv\theta + \frac{vu'}{l} + 2\alpha p\theta - \frac{\big(\eta(\theta)u'\big)'}{l^2} = 0, \tag{2.27}$$

$$-cu^2 + \frac{p'}{l} = 0, \tag{2.28}$$

$$\frac{v\theta'}{l} - \frac{\big(\lambda(\theta)\theta'\big)'}{l^2} = 0, \tag{2.29}$$

where we have introduced the notation $c = \mathcal{K}(0)/\rho_\sigma$ and we have $c > 0$ and $\alpha > 0$. Note that we have used the relation $\rho_\sigma/\rho = \theta$, which is valid for $\xi = 0$, since $U = 0$ on the symmetry line implies that $P = P_\sigma$ from (2.12) and (2.15). The boundary conditions are also given by (2.24) and (2.25) specialized to $\xi = 0$,

$$v(1) = v_\sigma, \qquad u(1) = u_\sigma, \qquad p(1) = p_\sigma, \qquad \theta(1) = \theta_\sigma, \tag{2.30}$$

$$v(0) = 0, \qquad u(0) = 0, \qquad \theta(0) = \theta_w, \tag{2.31}$$

where $v_\sigma < 0$, $u_\sigma = 1$, $p_\sigma < 0$, $\theta_\sigma = 1$, and $\theta_w > 0$. An explicit calculation also yields that $v_\sigma = -(\epsilon Re_\sigma/2)^{1/2}$, $p_\sigma = -2\epsilon/(\gamma+1)$ and $\theta_w = T_w/T_\sigma$ so that we may assume for convenience that $v_\sigma < -1$, $-1 < p_\sigma$ and $\theta_w < 1$, since in the thin layer approximation we have $\epsilon Re_\sigma \gg 1$ and $\epsilon \ll 1$ and since the body wall temperature is always lower than the shock temperature. Still note that these later assumptions are not strictly needed in the proof but simplify the analytic expression of various lower and upper bound introduced in the proof. The reduced transport coefficients η and λ are also functions of the reduced temperature θ only since $\rho_\sigma/\rho = \theta$, and we assume that these functions $\theta \to \eta(\theta)$ and $\theta \to \lambda(\theta)$ are smooth and such that $0 < \inf_{\mathbb{R}} \eta$, $\|\eta\|_{C^2(\mathbb{R})} < +\infty$, $0 < \inf_{\mathbb{R}} \lambda$, and $\|\lambda\|_{C^2(\mathbb{R})} < +\infty$.

Finally, we remark now that there are two boundary conditions for the reduced tangential velocity u and the reduced temperature θ which are governed by second-order equations, but that there are two boundary conditions for the reduced normal velocity v and one boundary condition for the reduced pressure p which are governed by first-order equations. The extra boundary condition for the reduced normal velocity v will thus be used to determine the eigenvalue l associated with the free boundary σ.

3. Existence of a solution. In this section, we prove the existence of a solution for the two point boundary value problem (2.26)–(2.31). The fundamental unknowns are v, u, p, θ, which are defined on $[0,1]$, and l. The method of proof is based on a priori estimates and on the Leray–Schauder topological degree theory. This method of proof is similar to the one used in [1] and [6] and differs from the one used by Weyl

[11] and McLeod and Serrin [9] for flat boundary layers which were based on shooting techniques.

3.1. A fixed point formulation. We introduce the Banach space $X = C^1[0,1] \times C^2[0,1] \times C^1[0,1] \times C^2[0,1] \times \mathbb{R}$ equipped with the norm

$$\|(v,u,p,\theta,l)\| = \max\left(\|v\|_{C^1[0,1]}, \|u\|_{C^2[0,1]}, \|p\|_{C^1[0,1]}, \|\theta\|_{C^2[0,1]}, |l|\right),$$

and the open set $\mathcal{O} = \{\, x = (v,u,p,\theta,l) \in X;\ l > 0 \,\}$, and we consider, for $\tau \in [0,1]$, the mapping K_τ from $\mathcal{O}$ to X defined by

$$K_\tau(v,u,p,\theta,l) = \Big(\mathcal{V},\mathcal{U},\mathcal{P},\mathcal{T}, l - \mathcal{V}(1) + v_\tau\Big), \tag{3.1}$$

where $\mathcal{V}$, $\mathcal{U}$, $\mathcal{P}$, and $\mathcal{T}$ are solutions of

$$(1-\tau)u + \tau\mathcal{U} + \frac{\mathcal{V}'}{l} = 0, \tag{3.2}$$

$$(1-\tau)\Big(\alpha(u^+)^2 + cu^+v\theta + \frac{vu'}{l} + 2\alpha p\theta\Big) + \tau\Big(\mathcal{U} - \frac{1}{2}\Big) - \frac{1}{l^2}\big(\eta_\tau(\theta)\mathcal{U}'\big)' = 0, \tag{3.3}$$

$$-(1-\tau)cu^2 + \frac{\mathcal{P}'}{l} = 0, \tag{3.4}$$

$$(1-\tau)\frac{v\theta'}{l} - \frac{1}{l^2}\big(\lambda_\tau(\theta)\mathcal{T}'\big)' = 0, \tag{3.5}$$

with the boundary conditions

$$\text{(a)}\ \mathcal{V}(0) = 0, \qquad \text{(b)}\ \mathcal{U}(0) = 0, \qquad \text{(c)}\ \mathcal{U}(1) = 1, \tag{3.6}$$

$$\text{(a)}\ \mathcal{P}(1) = p_\tau, \qquad \text{(b)}\ \mathcal{T}(0) = \theta_\tau, \qquad \text{(c)}\ \mathcal{T}(1) = 1, \tag{3.7}$$

where $u^+ = \max(0,u)$ and where

$$v_\tau = (1-\tau)v_\sigma - \tau, \qquad p_\tau = (1-\tau)p_\sigma - \tau, \qquad \theta_\tau = (1-\tau)\theta_w + \tau,$$

$$\eta_\tau = (1-\tau)\eta + \tau, \qquad \lambda_\tau = (1-\tau)\lambda + \tau.$$

Note that solutions of (2.26)–(2.31) such that $u \geq 0$ are fixed points of K_0, and the converse will be shown to be true in the following.

PROPOSITION 3.1. *The operator K_τ is well defined from $\mathcal{O}$ to X and for any closed bounded set $B \subset \mathcal{O}$, the mapping $(\tau,v,u,p,\theta,l) \to K_\tau(v,u,p,\theta,l)$ from $[0,1]\times B$ to X is compact.*

The proof of Proposition 3.1 relies of the following lemma.

LEMMA 3.2. *Let $\chi \in C^1[0,1]$ and $q \in C^0[0,1]$ be such that $\chi > 0$ and $q \geq 0$. Then for any $a, b \in \mathbb{R}$ and any $h \in C^0[0,1]$, the boundary value problem*

$$-(\chi\phi')' + q\phi = h, \qquad \phi(0) = a, \qquad \phi(1) = b,$$

has a unique solution $\phi \in C^2[0,1]$ *and* $\|\phi\|_{C^2[0,1]} \leq C\big(\|h\|_{C^0[0,1]} + |a| + |b|\big)$ *where the constant* C *only depends on* $\|\chi\|_{C^1[0,1]}$, $1/\inf_{[0,1]}\chi$ *and* $\|q\|_{C^0[0,1]}$. *Moreover, if* $\chi \in C^2[0,1]$, $q \in C^1[0,1]$, *and if* h *satisties a Lipschitz condition, i.e.,* $h \in \mathrm{Lip}[0,1]$, *then* $\phi'' \in \mathrm{Lip}[0,1]$ *and* $\|\phi''\|_{\mathrm{Lip}[0,1]} \leq \Gamma\big(\|h\|_{\mathrm{Lip}[0,1]} + |a| + |b|\big)$ *where the constant* Γ *only depends on* $\|\chi\|_{C^2[0,1]}$, $1/\inf_{[0,1]}\chi$ *and* $\|q\|_{C^1[0,1]}$ *and where we have defined* $\|h\|_{\mathrm{Lip}[0,1]} = \|h\|_{C^0[0,1]} + \sup\{\ |(h(x) - h(y)|/|x-y|,\ x,y \in [0,1],\ x \neq y\ \}$.

Sketch of the proof. Replacing ϕ and h by $\phi - \phi_0$ and $h + (\chi\phi_0')' - q\phi_0$, where $\phi_0(\zeta) = a + (b-a)\zeta$, it is easy to check that only the case $a = b = 0$ needs to be considered. In this situation, the weak formulation

$$\forall\psi \in H_0^1(0,1), \qquad \int_0^1 \chi(t)\phi'(t)\psi'(t)dt + \int_0^1 q(t)\phi(t)\psi(t)dt = \int_0^1 h(t)\psi(t)dt,$$

has a unique solution $\phi \in H_0^1$ from Poincaré's inequality and the Lax–Milgram Theorem. Using Poincaré's inequality $2\int_0^1 \phi^2(t)dt \leq \int_0^1 \phi'^2(t)dt$ also yields the estimate $(\inf_{[0,1]}\chi/2)\|\phi\|_{H_0^1(0,1)} \leq \|h\|_{L^2[0,1]}$. Moreover, we have $(\chi\phi')' = q\phi - h$ in the distribution sense so that $\chi\phi' \in H^1(0,1)$ and thus $\phi' \in H^1(0,1)$. This implies now that $\chi\phi'' = -\chi'\phi' + q\phi - h$ and, therefore, that $\|\phi\|_{H_0^2(0,1)} \leq C\|h\|_{L^2[0,1]}$. The two norm inequalities of the lemma are then easily obtained from these estimates and the relation $\chi\phi'' = -\chi'\phi' + q\phi - h$, making use of the inequalities $\|\psi\|_{C^0(0,1)} \leq C\|\psi\|_{H^1[0,1]}$ and $\|\psi\|_{L^2(0,1)} \leq C\|\psi\|_{C^0[0,1]}$.

Proof of Proposition 3.1. For a given $(v,u,p,\theta,l) \in \mathcal{O}$, we deduce from Lemma 3.2 that there exists a unique $\mathcal{U} \in C^2[0,1]$ such that (3.3), (3.6(b)), and (3.6(c)) hold and a unique $\mathcal{T} \in C^2[0,1]$ such that (3.5), (3.7(b)), and (3.7(c)) hold. The functions $\mathcal{V}$ and $\mathcal{P}$ can then be obtained by a simple integration and trivially $\big(\mathcal{V},\mathcal{U},\mathcal{P},\mathcal{T},l-\mathcal{V}(1)+v_T\big) \in X$. Moreover, for any closed bounded set $B \subset \mathcal{O}$, it is easy to check by using Lemma 3.2 and (3.2)–(3.7) that $\mathcal{V}'$, $\mathcal{U}''$, $\mathcal{P}'$, and $\mathcal{T}''$ are equi-Lipschitzian so that compactness is straightforward by using Ascoli's theorem.

Now we introduce the open bounded set Ω of X defined by

$$\begin{aligned}\Omega = \Big\{(v,u,p,\theta,l) \in X\ ;\ &\|v\|_{C^1[0,1]} < R,\ \|u\|_{C^2[0,1]} < R,\ \|p\|_{C^1[0,1]} < R,\\ &\|\theta\|_{C^2[0,1]} < R,\ 0 < A < l < B\Big\},\end{aligned} \tag{3.8}$$

where R, A, and B are positive constants, and the following proposition shows that the degree $d(I - K_\tau, \Omega, 0)$ can be defined for suitable R, A, and B.

PROPOSITION 3.3. *There exist constants* R, A, *and* B *such that*

$$\forall\tau \in [0,1] \quad (I - K_\tau)(\partial\Omega) \not\ni 0. \tag{3.9}$$

Proposition 3.3 relies on Lemmas 3.4 and 3.5 in which we derive strong estimates for fixed points of K_τ.

3.2. A priori estimates.

LEMMA 3.4. *There exist positive constants* m, A, *and* B *such that for any* $\tau \in [0,1]$ *and any fixed point* $(v,u,p,\theta,l) \in \mathcal{O}$ *of* K_τ *we have*

$$-m \leq v \leq 0, \qquad 0 \leq u \leq m, \qquad -m \leq p \leq 0, \tag{3.10}$$

$$0 \leq \theta \leq m, \qquad A < l < B. \tag{3.11}$$

Proof. Let $(v,u,p,\theta,l) \in \mathcal{O}$ be a fixed point of K_τ for some $\tau \in [0,1]$, i.e., be such that $v = \mathcal{V}$, $u = \mathcal{U}$, $p = \mathcal{P}$, $\theta = \mathcal{T}$, and $\mathcal{V}(1) = v_\tau$. By applying the maximum principle to the energy equations (3.5), we first deduce that $\theta_w \leq \theta_\tau \leq \theta \leq 1$. Hence θ is uniformly bounded by positive constants. From the normal momentum equation (3.4) we also deduce that p is nondecreasing so that $p(\zeta) \leq p_\tau \leq p_\sigma < 0$, and thus p is negative on $[0,1]$.

Assume now that the minimum of u over $[0,1]$ is negative. Then there exists a point $\zeta_0 \in (0,1)$ such that $u(\zeta_0) < 0$, $u'(\zeta_0) = 0$ and $u''(\zeta_0) \geq 0$ since $u(0) = 0$ and $u(1) = 1$. Then using the tangential momentum conservation equation (3.3) we obtain that

$$\eta_\tau\big(\theta(\zeta_0)\big)u''(\zeta_0) = (1-\tau)2\alpha p(\zeta_0)\theta(\zeta_0) + \tau u(\zeta_0) - \frac{\tau}{2},$$

which implies that $u''(\zeta_0) < 0$, a contradiction. Therefore, $u \geq 0$ over $[0,1]$ and u^+ can be replaced by u in (3.3). From (3.2) we now deduce that v is nonincreasing on $[0,1]$, and thus that $v_\sigma \leq v_\tau \leq v \leq 0$. We now want to estimate the maximum of u over $[0,1]$. Let us denote M as this maximum and assume that $M > 1$ since otherwise $u \leq M = 1$. Then there exists $\zeta_1 \in (0,1)$ such that $u(\zeta_1) = M$, $u'(\zeta_1) = 0$ and $u''(\zeta_1) \leq 0$. Using the tangential momentum conservation equation (3.3) we obtain that

$$(1-\tau)\alpha M^2 + \tau M \leq (1-\tau)\Big(-cMv(\zeta_1)\theta(\zeta_1) - 2\alpha p(\zeta_1)\theta(\zeta_1)\Big) + \frac{\tau}{2},$$

and since $M \geq 1$, $\theta_w \leq \theta \leq 1$ and $0 \leq -v \leq -v_\sigma$ we deduce that

$$M^2 \leq \left(\frac{c}{\alpha}\right)(-v_\sigma)M - 2p(\zeta_1).$$

However, by integrating the normal momentum equations (3.4), we obtain that

$$-p(\zeta_1) = -p_\tau + (1-\tau)cl\int_{\zeta_1}^1 u^2(t)dt \leq 1 + Mcl\int_0^1 u(t)dt,$$

since $-1 \leq p_\tau < 0$ and $0 \leq u \leq M$, and this implies that

$$-p(\zeta_1) \leq 1 + Mc(-v_\sigma), \tag{3.12}$$

since by integrating the mass conservation equation (3.2) and by using $v(0) = 0$ and $v(1) = v_\tau$ we obtain that

$$l\int_0^1 u(t)dt = -v_\tau \leq -v_\sigma. \tag{3.13}$$

Therefore, $M^2 \leq (2 + 1/\alpha)c(-v_\sigma)M + 2$ and thus $M^2 \leq (2+1/\alpha)(1+c)(-v_\sigma)M$ since $M \geq 1$ and $-v_\sigma \geq 1$ so that $M \leq (2+1/\alpha)(1+c)(-v_\sigma)$. We then deduce that p is bounded from (3.12), and letting $m > (2+1/\alpha)(1+c)^2(-v_\sigma)^2$ yields that $-m \leq v \leq 0$, $0 \leq u \leq m$, $-m \leq p \leq 0$, and $0 \leq \theta \leq m$.

We now estimate the eigenvalue l. From the relation (3.13) and since $u \leq m$ we first deduce that $1 \leq -v_\tau \leq lm$ so that $l \geq 1/m$, and we may choose $A < 1/m$. In order to obtain an upper bound for l, we then start with the tangential momentum equation (3.3) that we multiply by l^2 and that we write in the form

$$l^2\Big(-(1-\tau)2\alpha p\theta + \frac{\tau}{2}\Big) = (1-\tau)\Big((1+\alpha)l^2u^2 + cl^2uv\theta + l(vu)'\Big) + \tau l^2 u - \big(\eta_\tau(\theta)u'\big)' = 0,$$

making use of (3.2). Noting that the left-hand side coefficient $-(1-\tau)2\alpha p\theta + \tau/2$ is always larger than the positive constant $\delta = \min(-2\alpha\theta_w p_\sigma, \frac{1}{2})$ and that $cl^2uv\theta$ is nonpositive, we get that

$$l^2\delta \le (1-\tau)\Big((1+\alpha)l^2u^2 + l(vu)'\Big) + \tau l^2 u - \big(\eta_\tau(\theta)u'\big)'.$$

Integrating this inequality between r and s, where $0 \le r \le s \le 1$, and using the preceding a priori estimates then yields

$$l^2\delta(s-r) \le (1-\tau)\big((1+\alpha)l^2m\int_r^s u(t)\,dt - lv(r)u(r)\big) + \tau l^2\int_r^s u(t)\,dt - \int_r^s \big(\eta_\tau(\theta)u'\big)'(t)\,dt,$$

so that by using (3.13) we obtain

$$l^2\delta(s-r) \le lm(3+\alpha)(-v_\sigma) + \eta_\tau\big(\theta(r)\big)u'(r) - \eta_\tau\big(\theta(s)\big)u'(s).$$

Dividing this inequality by $\eta_\tau\big(\theta(r)\big)\eta_\tau\big(\theta(s)\big)$ we deduce that

$$\frac{l^2\delta(s-r)}{\overline{\eta}^2} \le \frac{lm(3+\alpha)(-v_\sigma)}{\underline{\eta}^2} + \frac{u'(r)}{\eta_\tau\big(\theta(s)\big)} - \frac{u'(s)}{\eta_\tau\big(\theta(r)\big)},$$

where $\underline{\eta} = \min(1, \inf_{\mathbb{R}}\eta)$, and $\overline{\eta} = \max(1, \sup_{\mathbb{R}}\eta)$, and integrating this inequality successively over $[0,s]$, with respect to r, and over $[0,1]$, with respect to s, and noting that

$$\int_0^1 \left(\int_0^s \left(\frac{u'(r)}{\eta_\tau\big(\theta(s)\big)} - \frac{u'(s)}{\eta_\tau\big(\theta(r)\big)}\right) dr\right) ds = \int_0^1 \frac{2u(s)-1}{\eta_\tau\big(\theta(s)\big)}\,ds \le \frac{2m}{\underline{\eta}},$$

we obtain that $l^2\delta \le l3m(3+\alpha)(-v_\sigma)\omega^2 + 12m\omega^2$, where $\omega = \overline{\eta}/\underline{\eta}$ and thus that $l \le 5m(3+\alpha)(-v_\sigma)\omega^2/\delta$. We may, therefore, choose $B > 5m(3+\alpha)(-v_\sigma)\omega^2/\delta$, and the proof is now complete. Note that the upper bound for the eigenvalue has been obtained because of the negative pressure term $2\alpha p\theta$ in the tangential momentum conservation equation. Remark also that the bounds A and B are such that $A < 2 < B$, which will be needed in the following.

From Lemma 3.4 and the relations (3.2)–(3.7) we now deduce the following result.

LEMMA 3.5. *There exist positive constants R, A, and B such that for any $\tau \in [0,1]$ and any fixed point $(v,u,p,\theta,l) \in \mathcal{O}$ of K_τ we have*

$$\|v\|_{C^1[0,1]} < R, \qquad \|u\|_{C^2[0,1]} < R, \qquad \|p\|_{C^1[0,1]} < R,$$

$$\|\theta\|_{C^2[0,1]} < R, \qquad A < l < B.$$

Proof. Keeping the notation of Lemma 3.4 we deduce from the relations (3.2) and (3.4) that $|v'| \le Bm$ and $|p'| \le Bcm^2$. From (3.3) and (3.2) we can also write that

$$\big(\eta_\tau(\theta)u'\big)' = l(1-\tau)(uv)' + \chi, \tag{3.14}$$

where $\chi = l^2(1-\tau)\big((1+\alpha)u^2 + cuv\theta + 2\alpha p\theta\big) + l^2\tau(u - \frac{1}{2})$. From Lemma 3.4 we know that χ is bounded independently of $(\tau, v, u, p, \theta, l)$, say $|\chi| \le K$. Integrating (3.14) between r and s, where $0 \leq r \leq s \leq 1$, and using the inequalities $|\chi| \leq K$ and (3.10), (3.11), we then obtain that

$$-(K + Bm^2) \leq \eta_\tau(\theta(s))u'(s) - \eta_\tau(\theta(r))u'(r) \leq K + Bm^2,$$

which is, therefore, valid for any $r, s \in [0,1]$. Dividing this inequality by $\eta_\tau(\theta(r))$, integrating with respect to r, between zero and 1, and dividing by $\int_0^1 (1/\eta_\tau(\theta))(r)\,dr$, we obtain that

$$-(K + Bm^2) \leq \eta_\tau(\theta(s))u'(s) \leq K + Bm^2 + \overline{\eta},$$

where $\overline{\eta} = \max(1, \sup_{\mathbf{R}} \eta)$, and dividing by $\eta_\tau(\theta(s))$ we get that $|u'| \leq (K+Bm^2+1)\omega$, where $\omega = \overline{\eta}/\underline{\eta}$ and $\underline{\eta} = \min(1, \inf_{\mathbf{R}} \eta)$. The same argument can now be applied to $|\theta'|$, and, therefore, $|u''|$ and $|\theta''|$ are also bounded independently of $(\tau, v, u, p, \theta, l)$ from (3.3) and (3.5) and the proof is complete.

3.3. Calculation of $d(I - K_\tau, \Omega, 0)$.

PROPOSITION 3.6. *Under the same hypotheses as Proposition* 3.3 *we have*

$$\forall \tau \in [0,1] \quad d(I - K_\tau, \Omega, 0) = -1.$$

Proof. From the homotopy invariance of the degree we know that $d(I-K_\tau, \Omega, 0) = d(I - K_1, \Omega, 0)$. But K_1 is a mapping depending only on l which reads

$$K_1(v, u, p, \theta, l) = (\mathcal{V}_1, \mathcal{U}_1, \mathcal{P}_1, \mathcal{T}_1, l - \mathcal{V}_1(1) - 1).$$

Introducing now the homotopy

$$H_\tau(v, u, p, \theta, l) = (\tau\mathcal{V}_1, \tau\mathcal{U}_1, \tau\mathcal{P}_1, \tau\mathcal{T}_1, l - \mathcal{V}_1(1) - 1),$$

we may easily check that the map $(\tau, v, u, p, \theta, l) \to H_\tau(v, u, p, \theta, l)$ from $[0,1] \times \overline{\Omega}$ to X is compact. Moreover, if (v, u, p, θ, l) is a fixed point of H_τ, then we have $v = \tau\mathcal{V}_1$, $u = \tau\mathcal{U}_1$, $p = \tau\mathcal{P}_1$, $\theta = \tau\mathcal{T}_1$ and $\mathcal{V}_1(1) = -1$. Thus $(\mathcal{V}_1, \mathcal{U}_1, \mathcal{P}_1, \mathcal{T}_1, l)$ is a fixed point of K_1, and we deduce from Proposition 3.3 that $\|\mathcal{V}_1\|_{C^1[0,1]} < R$, $\|\mathcal{U}_1\|_{C^2[0,1]} < R$, $\|\mathcal{P}_1\|_{C^1[0,1]} < R$, $\|\mathcal{T}_1\|_{C^2[0,1]} < R$, and $0 < A < l < B$. Hence $\|v\|_{C^1[0,1]} = \tau\|\mathcal{V}_1\|_{C^1[0,1]} < R$, $\|u\|_{C^2[0,1]} = \tau\|\mathcal{U}_1\|_{C^2[0,1]} < R$, $\|p\|_{C^1[0,1]} = \tau\|\mathcal{P}_1\|_{C^1[0,1]} < R$ and $\|\theta\|_{C^2[0,1]} = \tau\|\mathcal{T}_1\|_{C^2[0,1]} < R$. We have thus shown that

$$\forall \tau \in [0,1] \quad (I - H_\tau)(\partial\Omega) \not\ni 0,$$

so that $d(I - H_\tau, \Omega, 0)$ is well defined and $d(I - H_1, \Omega, 0) = d(I - H_0, \Omega, 0)$. Now since $H_1 = K_1$ we deduce that $d(I - K_0, \Omega, 0) = d(I - H_0, \Omega, 0)$, and since H_0 is a mapping which reads

$$H_0(v, u, p, \theta, l) = (0, 0, 0, 0, l - \mathcal{V}_1(1) - 1),$$

we deduce from the multiplicative property of the degree that

$$d(I - K_1, \Omega, 0) = d(\chi(l) + 1, (A, B), 0),$$

where we have defined $\chi(l) = \mathcal{V}_1(1)$. However, a straightforward calculation leads to

$$\mathcal{U}_1(\zeta) = \frac{1}{2}(1 - \mathrm{ch}(l\zeta)) + \frac{1 + \mathrm{ch}(l)}{2\mathrm{sh}(l)}(\mathrm{sh}(l\zeta)),$$

which implies by integration that

$$\mathcal{V}_1(\zeta) = -\frac{l\zeta}{2} + \frac{\mathrm{sh}(l\zeta)}{2} - \frac{1+\mathrm{ch}(l)}{2\mathrm{sh}(l)}(\mathrm{ch}(l\zeta) - 1),$$

and that $\mathcal{V}_1(1) = -l/2$ so that finally

$$\chi(l) = -\frac{l}{2}$$

(this simple result is the origin for choosing the constant $\frac{1}{2}$ as the limit of $-2\alpha p\theta$ in the homotopy path (3.3)) so that $d(I - K_1, \Omega, 0) = -1$ since we have seen in the proof of Lemma 3.4 that $A < 2 < B$.

We can now state the main result of this section.

THEOREM 3.7. *There exists a solution* $(v, u, p, \theta, l) \in \mathcal{O}$ *of the thin viscous shock layer problem* (2.26)–(2.31).

Remark. Due to the lack of knowledge on the linearized system deduced from (2.26)–(2.31), it is not known if such a solution is unique. A similar problem also arises for extending Oleinik's method of proof [10] to the system (2.20)–(2.25). More specifically, wellposedness and a priori estimates for linearized systems deduced from (2.20)–(2.25), with $\xi(\partial/\partial\xi)$ replaced by finite differences, would be needed in order to estimate the ξ derivatives of the solution components.

REFERENCES

[1] H. BERESTYCKI, B. NICOLAENKO, AND B. SCHEURER, *Travelling wave solutions to combustion models and their singular limits*, SIAM J. Math. Anal., 16 (1985), pp. 1207–1242.

[2] F. G. BLOTTNER, *Viscous shock layer at the stagnation point with nonequilibrium air chemistry*, AIAA J., 7 (1969), pp. 2281–2288.

[3] W. B. BUSH, *On the viscous hypersonic blunt body problem*, J. Fluid Mech., 20 (1964), pp. 353–367.

[4] A. H. CRAVEN AND L. A. PELETIER, *On the uniqueness of solutions of the Falkner–Skan equations*, Mathematika, 19 (1972), pp. 129–133.

[5] R. T. DAVIS, *Numerical solution of the hypersonic viscous shock layer equations*, AIAA J., 8 (1970), pp. 843–851.

[6] V. GIOVANGIGLI, *Non adiabatic plane laminar flames and their singular limits*, SIAM J. Math. Anal., 21(1990), pp. 1305–1325.

[7] H. HO AND R. F. PROBSTEIN, *The compressible viscous layer in rarefied hypersonic flow*, in Rarefied Gas Dynamics, L. Talbot, ed., Academic Press, New York, 1961, pp. 525–552.

[8] B. LABOUDIGUE, V. GIOVANGIGLI, AND S. CANDEL, *Numerical solution of a free-boundary problem in hypersonic flow theory: nonequilibrium viscous shock layers*, J. Comput. Phys., 102 (1992), pp. 297–309.

[9] J. B. MCLEOD AND J. SERRIN, *The existence of similar solutions for some laminar boundary layer problem*, Arch. Rational Mech. Anal., 31 (1968), pp. 288–303.

[10] O. A. OLEINIK, *On the system of boundary-layer equations for axisymmetric flows*, Soviet Math. Dokl., 8 (1967), pp. 921–925.

[11] H. WEYL, *On the differential equations of the simplest boundary layer problems*, Ann. of Math., 43 (1942), pp. 381–407.

SIAM J. MATH. ANAL.
Vol. 24, No. 3, pp. 583-602, May 1993

003

ON SOLUTIONS TO THE LINEAR BOLTZMANN EQUATION WITH EXTERNAL ELECTROMAGNETIC FORCE*

FRANTIŠEK CHVÁLA†, TOMMY GUSTAFSSON‡, AND ROLF PETTERSSON‡

Abstract. The authors consider the linear space-inhomogeneous Boltzmann equation in the full space $\mathbb{R}^3$ with external electromagnetic force. First, existence and uniqueness results about mild L^1-solutions are presented for soft and hard interactions, together with global boundedness in time for higher moments in the hard interaction case. Then, in the case of spatially homogeneous forces, mild $L^{1;q} \cap L^{p;q}$-solutions, $1 \leq p, q \leq \infty$, are constructed by an iteration procedure using an a priori estimate given by the corresponding homogeneous solution. Furthermore, some results about L^p-solutions are presented for space-inhomogeneous forces.

Key words. linear Boltzmann equation, external forces, electromagnetic force, mild solution, L^p-solution, $L^{1;q} \cap L^{p;q}$-solutions, higher moments, collision operator estimates

AMS subject classification. primary 76P05

Introduction. Consider a gas mixture consisting of two components, one formed by charged, the other by neutral, particles. Suppose that external electric and magnetic fields are imposed upon the mixture. We shall be concerned with the evolution of the distribution function $f(\mathbf{v}, \mathbf{x}, t)$ for the charged particles ($\mathbf{v} \in \mathbb{R}^3$, $\mathbf{x} \in \mathbb{R}^3$, and $t \in \mathbb{R}_+$ denoting the velocity, space, and time variables, respectively), in case the initial distribution $\varphi(\mathbf{v}, \mathbf{x})$ of the charged particles is given and the distribution function $F(\mathbf{v}, \mathbf{x}, t)$ for the neutral particles is known. Let each of the charged particles have mass m and charge q, and let the ratio of the mass of a charged particle to the mass of a neutral particle be κ. Denote by $\mathbf{\Gamma}(\mathbf{x}, t)$ and $\mathbf{\Omega}(\mathbf{x}, t)$ the (q/m)-multiples of the external electric and magnetic fields, respectively. Under some physically motivated hypotheses (cf. [Ce]) f is a solution of the following Cauchy problem for the linear Boltzmann kinetic equation:

$$\text{(0.1a)} \qquad \frac{\partial f}{\partial t} + \mathbf{v} \cdot \frac{\partial f}{\partial \mathbf{x}} + (\mathbf{\Gamma} + \mathbf{v} \times \mathbf{\Omega}) \cdot \frac{\partial f}{\partial \mathbf{v}} = Q[f, F] \quad \text{on } \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}_+,$$

$$\text{(0.1b)} \qquad f(\cdot, \cdot, 0) = \varphi \quad \text{on } \mathbb{R}^3 \times \mathbb{R}^3.$$

The differential operator on the left-hand side of (0.1a) is the streaming operator, and $Q[f, F]$ is the Bolztmann collision term. Recall that

$$Q[f, F](\mathbf{v}, \mathbf{x}, t) = \int (f'F'_* - fF_*) B(\theta, w) \, d\mathbf{v}_* \, d\theta \, d\varepsilon, \qquad \mathbf{v}, \mathbf{x} \in \mathbb{R}^3, \quad t \in \mathbb{R}_+,$$

where

$$f' = f(\mathbf{v}', \mathbf{x}, t), \quad F'_* = F(\mathbf{v}'_*, \mathbf{x}, t), \quad f = f(\mathbf{v}, \mathbf{x}, t), \quad F_* = F(\mathbf{v}_*, \mathbf{x}, t),$$
$$\mathbf{w} = \mathbf{v} - \mathbf{v}_*,$$
$$\mathbf{v}' = \mathbf{v} - \frac{2}{1+\kappa} (\mathbf{w} \cdot \mathbf{e}) \mathbf{e},$$
$$\mathbf{v}'_* = \mathbf{v}_* + \frac{2\kappa}{1+\kappa} (\mathbf{w} \cdot \mathbf{e}) \mathbf{e},$$
$$\mathbf{e} = (\sin\theta \cos\varepsilon, \sin\theta \sin\varepsilon, \cos\theta);$$

* Received by the editors June 25, 1990; accepted for publication (in revised form) August 31, 1992.

† Institute of Electrical Engineering, Czechoslovak Academy of Sciences, Praha, Czechoslovakia.

‡ Department of Mathematics, Chalmers University of Technology, University of Göteborg, Sweden.

$B:[0, \pi/2)\times(0,\infty)\to\mathbb{R}_+$ is closely related to the differential cross section; the range of integration is $\mathbb{R}^3\times[0,\pi/2)\times[0,2\pi)$. In accordance with the physical meaning F is measurable and nonnegative on $\mathbb{R}^3\times\mathbb{R}^3\times\mathbb{R}_+$.

Throughout this paper we use the convention $v=|\mathbf{v}|$, etc.

Supposing the collisions are governed by an inverse power-law interaction potential $U(\rho)=\text{const}\cdot\rho^{-(k-1)}(k>2)$ depending on the distance ρ of two colliding particles, the function B factorizes as follows [Ce]:

$$B(\theta, w)=w^{\gamma}\cdot b(\theta), \qquad 0\leqq\theta<\frac{\pi}{2}, \qquad w>0, \tag{0.2}$$

where $\gamma=(k-5)/(k-1)$ and b is a specified continuous nonnegative function on $[0,\pi/2)$ satisfying $b(\theta)=\mathcal{O}(\theta)$ when $\theta\to 0_+$, and $b(\theta)=\mathcal{O}((\pi/2-\theta)^{-(k+1)/(k-1)})$ when $\theta\to\pi/2_-$. Following [Gr], the interaction potential is called soft or hard acccordingly as $2<k\leqq 5$ or $k>5$, respectively. In the hard-sphere model of interaction the function B satisfies (0.2) with $\gamma=1$ and $b(\theta)=\sin\theta\cos\theta$. In the following we speak of soft or hard interactions for $-3<\gamma\leqq 0$ or $0\leqq\gamma\leqq 1$, respectively.

From the mathematical point of view it is considerably more convenient to redefine the function b in a left neighbourhood of $\pi/2$, so that for some $\alpha>-1$

$$b(\theta)=\mathcal{O}\left(\left(\frac{\pi}{2}-\theta\right)^{\alpha}\right), \quad \text{when } \theta\to\frac{\pi}{2}-,$$

a so-called angular cut-off. Physically it means ignoring part of the grazing collisions. For a detailed discussion of the cut-off technique, the reader is referred to [Ce], [TM], or [Ar].

Throughout this paper we consider B of the form (0.2), where $-3<\gamma\leqq 1$ and b is any nonnegative measurable function on $(0,\pi/2)$ satisfying

$$0<\int_0^{\pi/2} b(\theta)\,d\theta<\infty. \tag{0.3}$$

Particularly, (0.3) holds if b satisfies

$$0\leqq b(\theta)\leqq b_0\sin\theta(\cos\theta)^{\alpha}, \qquad 0\leqq\theta<\frac{\pi}{2}, b_0=\text{const}>0, \quad \alpha>-1. \tag{0.4}$$

As a consequence of (0.3) the collision term Q can be split as follows:

$$Q[f,F]=Kf-\nu\cdot f,$$

where K is the gain collision operator, defined by

$$Kf(\mathbf{v},\mathbf{x},t)=\int f'F'_*B(\theta,w)\,d\mathbf{v}_*\,d\theta\,d\varepsilon, \qquad \mathbf{v},\mathbf{x}\in\mathbb{R}^3, \quad t\in\mathbb{R}_+,$$

and ν is the collision frequency,

$$\nu(\mathbf{v},\mathbf{x},t)=2\pi\int F_*B(\theta,w)\,d\mathbf{v}_*\,d\theta, \qquad \mathbf{v},\mathbf{x}\in\mathbb{R}^3, \quad t\in\mathbb{R}_+ \tag{0.5}$$

(the ranges of integration for the former and the latter integrals are $\mathbb{R}^3\times[0,\pi/2)\times[0,2\pi)$ and $\mathbb{R}^3\times[0,\pi/2)$, respectively).

This leads to an alternative form of problem (0.1):

$$\frac{\partial f}{\partial t}+\mathbf{v}\cdot\frac{\partial f}{\partial \mathbf{x}}+(\mathbf{\Gamma}+\mathbf{v}\times\mathbf{\Omega})\cdot\frac{\partial f}{\partial \mathbf{v}}+\nu\cdot f=Kf \quad \text{on } \mathbb{R}^3\times\mathbb{R}^3\times\mathbb{R}_+, \tag{0.6a}$$

$$f(\cdot,\cdot,0)=\varphi \quad \text{on } \mathbb{R}^3\times\mathbb{R}^3. \tag{0.6b}$$

The aim of this paper is to present criteria for existence and uniqueness of a solution f of the problem (0.6), and to study boundedness of its moments, i.e., functions (of the time variable) $\|f_\sigma(\cdot,\cdot,t)\|$, where $f_\sigma(\mathbf{v},\mathbf{x},t)=(1+v^2)^{\sigma/2}f(\mathbf{v},\mathbf{x},t)$, $\sigma\geqq 0$ the norm $\|\cdot\|$ being taken in a convenient Lebesgue space.

The present paper, which is a shorter version of a preprint, below called [CGP], is based on earlier results by three authors [Ch], [Gu1], [P1], [P2]. The paper [Ch] studies properties of some operators related to the Boltzmann collision term. In [Gu1] L^p-estimates are given for solutions to the nonlinear, space-homogeneous Boltzmann equation. The papers [P1] and [P2] study solutions to the linear space-inhomogeneous Boltzmann equation in the periodic boundary case without external forces.

Section 1 of this paper presents the ordinary differential equations related to the external electromagnetic field. It also considers solutions to the linear Boltzmann equation in mild form in the case of external forces. In § 2 mild L^1-solutions are constructed by an iteration procedure both for hard and soft interactions. The main results in this section concern existence, uniqueness, and global boundedness in time for higher moments in the hard interaction case. Section 3 deals with mild $L^{1;q}\cap L^{p;q}$-solutions, $1\leqq p, q\leqq\infty$, constructed by an iteration procedure similar to that of § 2, when the external forces and the neutral particle distribution are spatially homogeneous. A necessary a priori estimate is obtained by introducing a new device, a spatially homogeneous solution, which is an upper bound for the L^q-norm with respect to the space variable of the solution from § 2. Section 4 gives some results on existence and uniqueness of L^p-solutions in the soft interaction case with spatially inhomogeneous external forces using boundedness of the gain term in the collision operator.

Remark 1. In the case of stationary external forces the results in §§ 2–4 concerning moments $\|(1+v^2)^{\sigma/2}f(\mathbf{v},\mathbf{x},t)\|$ can be generalized to moments corresponding to total energies, $\|(\phi(\mathbf{x})+v^2)^{\sigma/2}f(\mathbf{v},\mathbf{x},t)\|$. Here $\phi=\phi(\mathbf{x})$ is a potential for $\mathbf{\Gamma}$, $\mathbf{\Gamma}=-\frac{1}{2}\operatorname{grad}\phi$ with $\inf_{\mathbf{x}\in\mathbb{R}^3}\phi(\mathbf{x})=1$. For details on such results and their proofs, see § 5 in [CGP].

Remark 2. Notice that the linear Boltzmann equation, studied in this paper, is different from the linearized one, which is derived from the nonlinear equation. One essential difference with respect to methods is that the monotonicity property (e.g., $f_{n+1}\geqq f_n$), central to our approach, does not hold for the linearized equation.

1. Preliminaries. In connection with transforming problem (0.6) into a purely integral form, we shall be concerned with solution $\mathbf{v}=\mathbf{v}(t)=\mathbf{v}(\mathbf{u},\mathbf{y},t)$, $\mathbf{x}=\mathbf{x}(t)=\mathbf{x}(\mathbf{u},\mathbf{y},t)$ to the characteristic problem of the streaming operator:

$$\frac{d\mathbf{v}}{dt}=\mathbf{\Gamma}(\mathbf{x},t)+\mathbf{v}\times\mathbf{\Omega}(\mathbf{x},\mathrm{t}), \qquad \mathbf{v}(0)=\mathbf{u}, \tag{1.1a}$$

$$\frac{d\mathbf{x}}{dt}=\mathbf{v}, \qquad \mathbf{x}(0)=\mathbf{y}. \tag{1.1b}$$

In the rest of this paper we assume the following hypothesis.

Hypothesis 1.1.

(i) There exists a unique, locally absolutely continuous function on $\mathbb{R}_+$, satisfying (1.1) for almost every $t\in\mathbb{R}_+$.

(ii) The Jacobian of the transformation

$$(\mathbf{u},\mathbf{y})\mapsto(\mathbf{v}(\mathbf{u},\mathbf{y},t),\mathbf{x}(\mathbf{u},\mathbf{y},t)),\qquad \mathbf{u},\mathbf{y}\in\mathbb{R}^3,$$

is equal to 1 for every $t\in\mathbb{R}_+$.

PROPOSITION 1.2. *Hypothesis* 1.1 *is satisfied, if* $\mathbf{\Gamma}$ *and* $\mathbf{\Omega}$ *satisfy the following conditions of Carathéodory type*:

(i) *For every* $\mathbf{x}\in\mathbb{R}^3$ *the functions* $\mathbf{\Gamma}(\mathbf{x},\cdot)$, $\mathbf{\Omega}(\mathbf{x},\cdot)$ *are measurable on* $\mathbb{R}_+$;

(ii) *For every* $t\in\mathbb{R}_+$ *the functions* $\mathbf{\Gamma}(\cdot,t)$, $\mathbf{\Omega}(\cdot,t)$ *and all their first-order derivatives are continuous on* $\mathbb{R}^3$;

(iii) *There exists a function* $\rho_0\in L^1_{\mathrm{loc}}(\mathbb{R}_+)$ *such that*

$$|\mathbf{\Gamma}(\mathbf{x},t)|\leqq\rho_0(t),\qquad \mathbf{x}\in\mathbb{R}^3,\quad t\in\mathbb{R}_+,$$

$$|\mathbf{\Omega}(\mathbf{x},t)|\leqq\rho_0(t),\qquad \mathbf{x}\in\mathbb{R}^3,\quad t\in\mathbb{R}_+,$$

$$|\mathbf{\Gamma}(\mathbf{x}_1,t)-\mathbf{\Gamma}(\mathbf{x}_2,t)|\leqq\rho_0(t)\cdot|\mathbf{x}_1-\mathbf{x}_2|,\qquad \mathbf{x}_1,\mathbf{x}_2\in\mathbb{R}^3,\quad t\in\mathbb{R}_+,$$

$$|\mathbf{\Omega}(\mathbf{x}_1,t)-\mathbf{\Omega}(\mathbf{x}_2,t)|\leqq\rho_0(t)\cdot|x_1-x_2|,\qquad \mathbf{x}_1,\mathbf{x}_2\in\mathbb{R}^3,\quad t\in\mathbb{R}_+.$$

Proof. For an extensive discussion, see [CGP]; cf. also [Ku]. □

Remark. Cf. also Chapter XI, § 2, [GMP] about assumptions on the exterior Lorentz force $\mathbf{a}=\mathbf{\Gamma}(\mathbf{x},t)+\mathbf{v}\times\mathbf{\Omega}(\mathbf{x},t)$ to guarantee unique solvability of problem (1.1).

Notation. We employ a convention introduced in [KS]. For a function h defined on $\mathbb{R}^3\times\mathbb{R}^3\times\mathbb{R}_+$, the symbol $h^{\#}$ denotes the function specified on the same domain by

$$h^{\#}(\mathbf{u},\mathbf{y},t)=h(\mathbf{v}(t),\mathbf{x}(t),t)\qquad \mathbf{u},\mathbf{y}\in\mathbb{R}^3,\quad t\in\mathbb{R}_+,$$

where $\mathbf{v}(t),\mathbf{x}(t)$ is the solution to (1.1).

In connection with the problem (0.6) we shall also consider the following related problem:

$$\frac{\partial f}{\partial t}+\mathbf{v}\cdot\frac{\partial f}{\partial\mathbf{x}}+(\mathbf{\Gamma}+\mathbf{v}\times\mathbf{\Omega})\cdot\frac{\partial f}{\partial\mathbf{v}}+\nu\cdot f=g,\tag{1.2a}$$

$$f(\cdot,\cdot,0)=\varphi,\tag{1.2b}$$

where g is a given function defined on $\mathbb{R}^3\times\mathbb{R}^3\times\mathbb{R}_+$.

Notation. The Lebesgue measure in $\mathbb{R}^n$ is denoted by mes.

DEFINITION. Let Hypothesis 1.1 be satisfied, and let

$$\nu^{\#}(\mathbf{u},\mathbf{y},\cdot)\in L^1_{\mathrm{loc}}(\mathbb{R}_+),\ g^{\#}(\mathbf{u},\mathbf{y},\cdot)\in L^1_{\mathrm{loc}}(\mathbb{R}_+),\qquad \mathbf{u},\mathbf{y}\in\mathbb{R}^3.\tag{1.3}$$

A real-valued function f is called a *mild solution* of the problem (1.2) if there exists $M\subset\mathbb{R}^3\times\mathbb{R}^3$, mes $M=0$, such that $f^{\#}$ is defined on $(\mathbb{R}^3\times\mathbb{R}^3\backslash M)\times\mathbb{R}_+$ and satisfies the relation

$$\begin{gathered}f^{\#}(\mathbf{u},\mathbf{y},t)=\varphi(\mathbf{u},\mathbf{y})+\int_0^t g^{\#}(\mathbf{u},\mathbf{y},s)\,ds-\int_0^t\nu^{\#}(\mathbf{u},\mathbf{y},s)f^{\#}(\mathbf{u},\mathbf{y},s)\,ds,\\ (\mathbf{u},\mathbf{y})\in\mathbb{R}^3\times\mathbb{R}^3\backslash M,\qquad t\in\mathbb{R}_+.\end{gathered}\tag{1.4}$$

Another formulation of (1.2) is the *exponential form*

$$\begin{gathered}f^{\#}(\mathbf{u},\mathbf{y},t)=\varphi(\mathbf{u},\mathbf{y})\exp\left(-\int_0^t\nu^{\#}(\mathbf{u},\mathbf{y},s)\,ds\right)\\ +\int_0^t g^{\#}(\mathbf{u},\mathbf{y},r)\exp\left(-\int_r^t\nu^{\#}(\mathbf{u},\mathbf{y},s)\,ds\right)dr,\\ (\mathbf{u},\mathbf{y})\in\mathbb{R}^3\times\mathbb{R}^3\backslash M,\quad t\in\mathbb{R}_+.\end{gathered}\tag{1.5}$$

By straightforward calculations we obtain the following lemma on the connection between (1.4) and (1.5).

LEMMA 1.3. *Let* (1.3) *be satisfied. Then f is a mild solution of problem* (1.2) *if and only if the exponential form* (1.5) *holds.*

Remark. For an extensive discussion see [PL] (cf. also [P1]).

DEFINITION. The notion of a mild solution of the problem (0.6) is defined in a similar manner, viz. as a function f such that $(Kf)^{\#}(\mathbf{u},\mathbf{y},\cdot)\in L^1_{\mathrm{loc}}(\mathbb{R}_+)$, $(\mathbf{u},\mathbf{y})\in\mathbb{R}^3\times\mathbb{R}^3\backslash M$, and satisfying (1.4) with g replaced by Kf.

Remark. In our paper mild solutions play a central role. They are mathematically easy to use, and also physically natural. (For instance the iterate f_{n+1}, constructed for generating a solution, represents the distribution of particles undergone at most n collisions.) In dealing with the (different) nonlinear and linearized Boltzmann equations, other solution concepts are often used, e.g., the renormalized solutions; cf. [PL]. In the present situation, however, such devices seem to introduce unnecessarily technical complications without adding new insights.

2. L^1-solutions. In this section a mild L^1-solution of the problem (0.6) is constructed as a limit of a monotone sequence of suitably defined iterates. The construction was used in [P1], [P2] to treat a (periodic) boundary value problem for the linear Boltzmann equation without external forces; now it is applied to the Cauchy problem for the full linear Boltzmann equation. Under general conditions we prove existence of a solution together with mass conservation, and uniqueness. We also prove that higher moments of the solution exist (if they do so initially) for both soft and hard interactions, and that they are globally bounded in the hard interaction case.

Notation. Let $1\leqq p\leqq\infty$. The positive cone of the space L^p is denoted by L^p_+. The weighted L^p-space with the weight function $(1+v^2)^{\sigma/2}$, $\sigma\geqq 0$ is denoted by L^p_σ and the corresponding norm by $\|\cdot\|_{p,\sigma}$. Set $\|\cdot\|_{1\cap p}=\|\cdot\|_1+\|\cdot\|_p$ and $\|\cdot\|_{1\cap p,\sigma}=\|\cdot\|_{1,\sigma}+\|\cdot\|_{p,\sigma}$.

Note that the collision frequency is well defined; cf. (0.2), (0.3), and (0.5), whenever

$$\text{(2.1a)}\qquad F(\cdot,\mathbf{x},t)\in L^1_\gamma(\mathbb{R}^3),\quad \mathbf{x}\in\mathbb{R}^3,\quad t\in\mathbb{R}_+\quad \text{if } 0\leqq\gamma\leqq 1,$$

$$\text{(2.1b)}\qquad F(\cdot,\mathbf{x},t)\in L^1(\mathbb{R}^3)\cap L^\infty(\mathbb{R}^3),\quad \mathbf{x}\in\mathbb{R}^3,\quad t\in\mathbb{R}_+\quad \text{if } -3<\gamma<0;$$

indeed, in the latter case there holds for every $\mathbf{v},\mathbf{x}\in\mathbb{R}^3$, $t\in\mathbb{R}_+$:

$$\int_{\mathbb{R}^3} w^\gamma F(\mathbf{v}_*,\mathbf{x},t)\,d\mathbf{v}_*\leqq \operatorname*{ess\text{-}sup}_{v_*\in\mathbb{R}^3} F(\mathbf{v}_*,\mathbf{x},t)\cdot\int_{|\mathbf{v}_*-\mathbf{v}|<1} w^\gamma\,d\mathbf{v}_*+\int_{\mathbb{R}^3} F(\mathbf{v}_*,\mathbf{x},t)\,d\mathbf{v}_*.$$

The following relation is valid whenever one of the integrals exists [Ce]:

$$\text{(2.2)}\qquad \int_{\mathbb{R}^3}(Kf)(\mathbf{v},\mathbf{x},t)\,d\mathbf{v}=\int_{\mathbb{R}^3}\nu(\mathbf{v},\mathbf{x},t)f(\mathbf{v},\mathbf{x},t)\,d\mathbf{v},\qquad \mathbf{x}\in\mathbb{R}^3,\quad t\in\mathbb{R}_+.$$

LEMMA 2.1. *Let F satisfy* (2.1), *and let $\{f_n\}_{n=0}^\infty$ be a sequence of functions defined on $\mathbb{R}^3\times\mathbb{R}^3\times\mathbb{R}_+$ recursively by*

$$f_0^{\#}(\mathbf{u},\mathbf{y},t)=0,\quad \mathbf{u},\mathbf{y}\in\mathbb{R}^3,\quad t\in\mathbb{R}_+,$$

$$\text{(2.3)}\qquad \begin{aligned} f_{n+1}^{\#}(\mathbf{u},\mathbf{y},t)&=\varphi(\mathbf{u},\mathbf{y})\exp\left(-\int_0^t \nu^{\#}(\mathbf{u},\mathbf{y},s)\,ds\right)+\int_0^t (Kf_n)^{\#}(\mathbf{u},\mathbf{y},r)\\ &\quad\cdot\exp\left(-\int_r^t \nu^{\#}(\mathbf{u},\mathbf{y},s)\,ds\right)dr,\quad \mathbf{u},\mathbf{y}\in\mathbb{R}^3,\quad t\in\mathbb{R}_+,\quad n=0,1,\ldots.\end{aligned}$$

Then $\{f_n\}_{n=0}^\infty$ is a nondecreasing sequence if φ is a nonnegative measurable function on $\mathbb{R}^3\times\mathbb{R}^3$.

Proof. By induction, cf. [P1]. □

LEMMA 2.2. *If the hypothesis of Lemma* 2.1 *is satisfied and* $\varphi\in L^1_+(\mathbb{R}^3\times\mathbb{R}^3)$, *then for every* $n=0,1,\dots,$

$$\|f_n(\cdot,\cdot,t)\|_1\leqq\|\varphi\|_1,\qquad t\in\mathbb{R}_+.$$

Proof. By (2.3) and Lemma 1.3 the following mild form holds:

$$(2.4)\quad f_n^{\#}(\cdot,\cdot t)+\int_0^t\nu^{\#}(\cdot,\cdot,s)f_n^{\#}(\cdot,\cdot s)\,ds=\varphi+\int_0^t(Kf_{n-1})^{\#}(\cdot,\cdot,s)\,ds,\qquad t\in\mathbb{R}_+,$$

where (by induction) all terms belong to $L^1_+(\mathbb{R}^3\times\mathbb{R}^3)$, $t\in\mathbb{R}_+$.

By integrating (2.4) we get, in view of relation (2.2), Hypothesis 1.1(ii), and Lemma 2.1,

$$\|f_n(\cdot,\cdot t)\|_1-\|\varphi\|_1=\int_0^t\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}[K(f_{n-1}-f_n)](\mathbf{v},\mathbf{x},s)\,d\mathbf{v}\,d\mathbf{x}\,ds\leqq 0,\qquad t\in\mathbb{R}_+.\quad\square$$

THEOREM 2.3. *Let* (2.1) *hold and* $\varphi\in L^1_+(\mathbb{R}^3\times\mathbb{R}^3)$; *then there exists a mild solution* f *of the problem* (0.6) *such that* $f(\cdot,\cdot,t)\in L^1_+(\mathbb{R}^3\times\mathbb{R}^3), t\in\mathbb{R}_+$, *and*

$$(2.5)\qquad \|f(\cdot,\cdot,t)\|_1\leqq\|\varphi\|_1,\qquad t\in\mathbb{R}_+.$$

Proof. Set $f(\mathbf{v},\mathbf{x},t)=\lim_{n\to\infty}f_n(\mathbf{v},\mathbf{x},t)$, $(\mathbf{v},\mathbf{x})\in\mathbb{R}^3\times\mathbb{R}^3$, $t\in\mathbb{R}_+$, where $\{f_n\}_0^\infty$ is the nondecreasing sequence defined in Lemma 2.1. Then by Levi's monotone convergence theorem f is measurable, and, using also Lemma 2.2, we have

$$\|f(\cdot,\cdot,t)\|_1=\lim_{n\to\infty}\|f_n(\cdot,\cdot,t)\|_1\leqq\|\varphi\|_1,\qquad t\in\mathbb{R}_+.$$

Moreover, letting $n\to\infty$ in (2.4), we see that there exists $M\subset\mathbb{R}^3\times\mathbb{R}^3$ such that mes $M=0$, and

$$f^{\#}(\mathbf{u},\mathbf{y},t)=\varphi(\mathbf{u},\mathbf{y})+\int_0^t(Kf)^{\#}(\mathbf{u},\mathbf{y},s)\,ds-\int_0^t\nu^{\#}(\mathbf{u},\mathbf{y},s)f^{\#}(\mathbf{u},\mathbf{y},s)\,ds,$$

$$(\mathbf{u},\mathbf{y})\in\mathbb{R}^3\times\mathbb{R}^3\backslash M,\qquad t\in\mathbb{R}_+.\quad\square$$

Remark. For $-3<\gamma\leqq0$ it follows by (2.1), (2.2), and the proof of Theorem 2.3 that equality holds in relation (2.5). For $0<\gamma<1$, see Theorem 2.5 below.

Notation. Given $\sigma\geqq0$, $R>1$, we define a function $h_{\sigma,R}:\mathbb{R}_+\to\mathbb{R}_+$ as follows:

$$(2.6)\qquad \begin{aligned} h_{\sigma,R}(u)&=(1+u^2)^{\sigma/2}\quad\text{for } 0\leqq u\leqq R-1,\\ h_{\sigma,R}(u)&=\text{constant}\quad\text{for } R\leqq u<\infty,\end{aligned}$$

and $dh_{\sigma,R}/du$ is continuous on $\mathbb{R}_+$ and decreasing on $(R-1,R)$. Denote for measurable functions $g:\mathbb{R}^3\to\mathbb{R}$,

$$\|g\|_{1,\sigma,R}=\|h_{\sigma,R}g\|_1.$$

Notation.

$$\tilde{\Gamma}(t)=\operatorname*{ess-sup}_{\mathbf{x}\in\mathbb{R}^3}|\mathbf{\Gamma}(\mathbf{x},t)|,\qquad t\in\mathbb{R}_+,\quad\text{and}$$

$$\tilde{\Gamma}_0=\operatorname*{ess-sup}_{\mathbf{x}\in\mathbb{R}^3,t\in\mathbb{R}_+}|\mathbf{\Gamma}(\mathbf{x},t)|.$$

If $\mathbf{v}(t),\mathbf{x}(t)$ is the solution of (1.1), then there holds

$$(2.7)\qquad \begin{aligned} &\frac{d}{dt}(1+v^2(t))^{\sigma/2}=\sigma(1+v^2(t))^{\sigma/2-1}\mathbf{v}(t)\cdot\mathbf{\Gamma}(\mathbf{x}(t),t),\qquad t\in\mathbb{R}_+,\\ &\left|\frac{d}{dt}h_{\sigma,R}(v(t))\right|\leqq\sigma\tilde{\Gamma}(t)(1+v^2(t))^{\sigma/2-1}v(t),\qquad t\in\mathbb{R}_+.\end{aligned}$$

These elementary relations are derived noticing that $d\mathbf{v}(t)/dt = \mathbf{\Gamma}(\mathbf{x}(t), t) + \mathbf{v}(t) \times \mathbf{\Omega}(\mathbf{x}(t), t)$ and $\mathbf{v} \cdot (\mathbf{v} \times \mathbf{\Omega}) = 0$.

The following proposition gives local boundedness in time for moments of our mild solution.

THEOREM 2.4. *Let* $\tilde{\Gamma} \in L^1_{\mathrm{loc}}(\mathbb{R}_+)$, $-3 < \gamma < 1$, $\sigma_0 \geqq 2$. *When* $-1 \leqq \gamma$, *let* (2.1) *hold together with* $\text{ess-sup}_{\mathbf{x}\in\mathbb{R}^3, t\in\mathbb{R}_+} \|F(\cdot, \mathbf{x}, t)\|_{1,\sigma_0+\gamma} < \infty$; *otherwise, let* $\text{ess-sup}_{\mathbf{x}\in\mathbb{R}^3, t\in\mathbb{R}_+} \|F(\cdot, \mathbf{x}, t)\|_{1\cap\infty,\sigma_0-1} < \infty$. *If* $\varphi \in L^1_{\sigma,+}(\mathbb{R}^3 \times \mathbb{R}^3)$ *for some* σ *such that* $0 < \sigma \leqq \sigma_0$, *then the mild solution* f (*given in Theorem* 2.3) *satisfies*

$$\|f(\cdot, \cdot, t)\|_{1,\sigma} \leqq \|\varphi\|_{1,\sigma} \exp(\psi(t)), \qquad t \in \mathbb{R}_+,$$

where $\psi(t) = \sigma(\int_0^t \tilde{\Gamma}(s)\, ds + ct)$ *and* c *is a positive constant.*

Proof. Let $\{f_n\}_{n=0}^{\infty}$ be the sequence of iterates (defined in Lemma 2.1) with limit f. Using (2.4) we have, for $t \in \mathbb{R}_+$,

$$\frac{\partial}{\partial t}(f_n^{\#}) = (Kf_{n-1})^{\#} - (\nu f_n)^{\#} \leqq (Kf_n)^{\#} - (\nu f_n)^{\#} = Q[f_n, F]^{\#}.$$

By multiplying with $h^{\#}_{\sigma,R}$ and integrating we get

$$h^{\#}_{\sigma,R} f_n^{\#} \leqq h_{\sigma,R}\varphi + \int_0^t \frac{\partial h^{\#}_{\sigma,R}}{\partial s} \cdot f_n^{\#}\, ds + \int_0^t h^{\#}_{\sigma,R} Q^{\#}[f_n, F]\, ds.$$

Further integration gives (after a change of variables)

$$\iint h_{\sigma,R} f_n\, d\mathbf{x}\, d\mathbf{v} \leqq \|\varphi\|_{1,\sigma} + \sigma \int_0^t \tilde{\Gamma}(s) \|f_n(\cdot, \cdot, s)\|_{1,\max(0,\sigma-1)}\, \mathrm{ds}$$

$$+ \int_0^t \iiiint\!\!\int [h_{\sigma,R}(v') - h_{\sigma,R}(v)]$$

$$\cdot f_n F_* B\, d\theta\, d\varepsilon\, d\mathbf{v}_*\, d\mathbf{x}\, d\mathbf{v}\, ds.$$

Here

$$h_{\sigma,R}(v') - h_{\sigma,R}(v) \leqq \sigma C_1 \cos\theta w (1 + v^2)^{(\sigma-2)/2} (1 + v_*)^{\max(1,\sigma-1)}$$

for some positive constant C_1 (see [P2] if $v' > v$). Thus (with a constant C_2)

$$\iint h_{\sigma,R} f_n\, d\mathbf{x}\, d\mathbf{v} \leqq \|\varphi\|_{1,\sigma} + \sigma \int_0^t \tilde{\Gamma}(s) \|f_n(\cdot, \cdot. s)\|_{1,\max(0,\sigma-1)}\, ds$$

$$+ \sigma C_2 \int_0^t \|f_n(\cdot, \cdot s)\|_{1,\max(0,\tilde{\sigma})}\, ds$$

$$\leqq \|\varphi\|_{1,\sigma} + \sigma \int_0^t (\tilde{\Gamma}(s) + C_2) \|f_n(\cdot, \cdot, s)\|_{1,\sigma-\delta}\, ds,$$

where $\tilde{\sigma} = \sigma + \gamma - 1$ if $\gamma \geqq -1$, $\tilde{\sigma} = \sigma - 2$ if $-3 < \gamma < -1$, and the last member of these inequalities is independent of R and valid for some $\delta > 0$.

To obtain the first inequality in the case $-3 < \gamma < -1$ we split the integral with respect to $\mathbf{v}_*$ into two parts, one corresponding to $|\mathbf{v} - \mathbf{v}_*| \leqq 1$ and the other to $|\mathbf{v} - \mathbf{v}_*| \geqq 1$, and thus

$$\int_{\mathbb{R}^3} (1 + v_*^2)^{(\sigma-1)/2} F(\mathbf{v}_*, \mathbf{x}, s) w^{1+\gamma}\, d\mathbf{v}_* \leqq C \cdot \underset{\mathbf{x}\in\mathbb{R}^3, t\in\mathbb{R}_+}{\text{ess-sup}} \|F(\cdot, \mathbf{x}, t)\|_{1\cap\infty,\sigma_0-1}.$$

Next let $R \to \infty$. Then $h_{\sigma,R}(v) \nearrow (1+v^2)^{\sigma/2}$, and

$$\|f_n(\cdot,\cdot,t)\|_{1,\sigma} \leqq \|\varphi\|_{1,\sigma} + \sigma \int_0^t (\tilde{\Gamma}(s)+C_2)\|f_n(\cdot,\cdot,s)\|_{1,\sigma-\delta}\,ds.$$

Using Gronwall's lemma, we obtain

$$\|f_n(\cdot,\cdot,t)\|_{1,\sigma} \leqq \|\varphi\|_{1,\sigma} \exp(\psi(t)),$$

where $\psi(t) = \sigma(\int_0^t \tilde{\Gamma}(s)\,ds + C_2 t)$, $t \in \mathbb{R}_+$. Letting $n \to \infty$, and noticing $f_n \nearrow f$, the theorem follows. □

Now, employing Theorem 2.4 we find the following theorem about mass conservation and uniqueness for the mild solution from Theorem 2.3.

THEOREM 2.5. *Let* $\tilde{\Gamma} \in L^1_{\text{loc}}(\mathbb{R}_+)$, $-3<\gamma<1$, *and let* (2.1) *hold together with* $\text{ess-sup}_{\mathbf{x}\in\mathbb{R}^3, t\in\mathbb{R}_+} \|F(\cdot,\mathbf{x},t)\|_{1,2+\max(-1,\gamma)} < \infty$. *If* $\varphi \in L^1_+(\mathbb{R}^3\times\mathbb{R}^3)$, *then there exists a nonnegative mild solution f to the problem* (0.6), *such that*

$$\|f(\cdot,\cdot,t)\|_1 = \|\varphi\|_1, \qquad t\in\mathbb{R}_+. \tag{2.8}$$

Moreover, the solution f is unique in the following sense: if $\tilde{f} = \tilde{f}(\cdot,\cdot,t) \in L^1_+(\mathbb{R}^3\times\mathbb{R}^3)$, $t\in\mathbb{R}_+$, *is a mild solution of* (0.6) *such that* $\|\tilde{f}(\cdot,\cdot,t)\|_1 \leqq \|\varphi\|_1$, $t\in\mathbb{R}_+$, *then* $\tilde{f}(\cdot,\cdot,t) = f(\cdot,\cdot,t)$ *almost everywhere in* $\mathbb{R}^3\times\mathbb{R}^3$, $t\in\mathbb{R}_+$.

Proof. Consider the case $0 \leqq \gamma < 1$. Let $\varphi_R = \varphi\cdot\chi_R$, where $\chi_R(\mathbf{v},\mathbf{x}) = 1$ if $|\mathbf{v}| \leqq R$ and zero otherwise. Let f and f_R be the solutions obtained in Theorem 2.3 with initial data φ and φ_R, respectively. Since $\varphi_R \in L^1_{\sigma,+}(\mathbb{R}^3\times\mathbb{R}^3)$, $t\in\mathbb{R}_+$, for all $\sigma \geqq 0$, it follows by Theorem 2.4 that $f_R(\cdot,\cdot,t) \in L^1_{\sigma,+}(\mathbb{R}^3\times\mathbb{R}^3)$, $t\in\mathbb{R}_+$, for $0\leqq\sigma\leqq 2$. Choose especially $\sigma=\gamma$. Thus both the right-hand side and the left-hand side of (2.2) with f replaced by f_R exist and are equal. Now integrate the appropriate modification of (2.4), and let $n\to\infty$. We conclude that

$$\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} f_R(\mathbf{v},\mathbf{x},t)\,d\mathbf{v}\,d\mathbf{x} = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \varphi_R(\mathbf{v},\mathbf{x})\,d\mathbf{v}\,d\mathbf{x}, \qquad t\in\mathbb{R}_+. \tag{2.9}$$

Next, let $\{f_n\}$ and $\{f_{nR}\}$ be the sequences from Theorem 2.3 with limits f and f_R, respectively. Then, by induction, $f_{nR} \leqq f_n$, and hence $f_R \leqq f$. So by Theorem 2.3 and (2.9) we have

$$\begin{aligned}\|f(\cdot,\cdot,t)-f_R(\cdot,\cdot,t)\|_1 &= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} f(\mathbf{v},\mathbf{x},t)\,d\mathbf{v}\,d\mathbf{x} - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} f_R(\mathbf{v},\mathbf{x},t)\,d\mathbf{v}\,d\mathbf{x}\\ &\leqq \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \varphi(\mathbf{v},\mathbf{x})\,d\mathbf{v}\;d\mathbf{x} - \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \varphi_R(\mathbf{v},\mathbf{x})\,d\mathbf{v}\,d\mathbf{x}\\ &= \|\varphi(\cdot,\cdot)-\varphi_R(\cdot,\cdot)\|_1, \quad t\in\mathbb{R}_+.\end{aligned}$$

Now let $R\to\infty$. We conclude that $f_R(\cdot,\cdot,t)\to f(\cdot,\cdot,t)$ in $L^1(\mathbb{R}^3\times\mathbb{R}^3)$ when $R\to\infty$, $t\in\mathbb{R}_+$. Hence (2.8) follows.

For the case $-3<\gamma<0$, see the remark after Theorem 2.3.

It remains to prove uniqueness. Let $\tilde{f}$ be any mild nonnegative solution of (0.6). Then by Lemma 1.3 $\tilde{f}$ satisfies also the exponential form (1.5) with $g = K\tilde{f}$. Hence

$$\tilde{f}^*(\cdot,\cdot,t) - f^*_{n+1}(\cdot,\cdot,t) = \int_0^t K(\tilde{f}-f_n)^*(\cdot,\cdot,r)\cdot\exp\left(-\int_r^t \nu^*(\cdot,\cdot,s)\,ds\right)dr,$$

and, by induction we find that $f_n \leqq \tilde{f}$, $n = 0, 1, \dots$. So $f(\cdot, \cdot, t) - \tilde{f}(\cdot, \cdot, t) \leqq 0$, $t \in \mathbb{R}_+$. But by (2.8),

$$\|f(\cdot, \cdot, t)\|_1 = \|\varphi\|_1 \geqq \|\tilde{f}(\cdot, \cdot, t)\|_1, \qquad t \in \mathbb{R}_+.$$

Thus, $f(\cdot, \cdot, t) = \tilde{f}(\cdot, \cdot, t)$ almost everywhere in $\mathbb{R}^3 \times \mathbb{R}^3$, $t \in \mathbb{R}_+$. □

We have arrived at the main result of this section, global boundedness in time for moments of the mild solution in the hard interaction case.

THEOREM 2.6. *Let* $\tilde{\Gamma}_0 < \infty$, $0 \leqq \gamma < 1$, *and let* ess-sup$_{\mathbf{x} \in \mathbb{R}^3, t \in \mathbb{R}_+}$ $\|F(\cdot, \mathbf{x}, t)\|_{1, \sigma_0 + \gamma} < \infty$ *for some* $\sigma_0 \geqq 2$, *and* ess-inf$_{\mathbf{x} \in \mathbb{R}^3, t \in \mathbb{R}_+}$ $\|F(\cdot, \mathbf{x}, t)\|_1 > 0$. *If* $\varphi \in L^1_{\sigma,+}(\mathbb{R}^3 \times \mathbb{R}^3)$ *for some* σ *such that* $0 \leqq \sigma \leqq \sigma_0$, *then there exists a unique nonnegative mild solution f of the problem* (0.6) *such that*

$$\|f(\cdot, \cdot, t)\|_1 = \|\varphi\|_1, \qquad t \in \mathbb{R}_+,$$

$$\|f(\cdot, \cdot, t)\|_{1,\sigma} \leqq C \|\varphi\|_{1,\sigma}, \qquad t \in \mathbb{R}_+,$$

where C is a constant not depending on φ.

Proof. Let φ_R and f_R be as in the proof of Theorem 2.5. Then the solution f_R satisfies

$$\frac{\partial f_R^\#(\mathbf{u}, \mathbf{y}, t)}{\partial t} = Q[f_R, F]^\#(\mathbf{u}, \mathbf{y}, t), (\mathbf{u}, \mathbf{y}) \in \mathbb{R}^3 \times \mathbb{R}^3 \backslash M, \qquad t \in \mathbb{R}_+ \backslash M_{\mathbf{u},\mathbf{y}},$$

with mes $M =$ mes $M_{\mathbf{u},\mathbf{y}} = 0$. By multiplying by $(1 + v(t)^2)^{\sigma/2}$ and integrating, we get (after some changes of variables)

$$\begin{aligned} &\int_{\mathbb{R}^3} \int_{\mathbb{R}^3} (1 + v^2)^{\sigma/2} f_R(\mathbf{v}, \mathbf{x}, t) \, d\mathbf{v} \, d\mathbf{x} \\ &= \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} (1 + v^2)^{\sigma/2} \varphi(\mathbf{v}, \mathbf{x}) \, d\mathbf{v} \, d\mathbf{x} \\ &\quad + \sigma \int_0^t \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} (1 + v^2)^{(\sigma-2)/2} (\mathbf{v} \cdot \mathbf{\Gamma}(\mathbf{x}, s)) f_R(\mathbf{v}, \mathbf{x}, s) \, d\mathbf{v} \, d\mathbf{x} \, ds + I(t), \qquad t \in \mathbb{R}_+. \end{aligned} \tag{2.10}$$

Using $f_R \in L^1_{\sigma_0 + \gamma}$ (cf. Theorem 2.4),

$$\begin{aligned} I(t) = \int_0^t \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \int_0^{2\pi} \int_0^{\pi/2} &[(1 + (v')^2)^{\sigma/2} - (1 + v^2)^{\sigma/2}] \\ &\cdot f_R(\mathbf{v}, \mathbf{x}, s) F(\mathbf{v}_*, \mathbf{x}, s) B(\theta, w) \, d\theta \, d\varepsilon \, d\mathbf{v}_* \, d\mathbf{v} \, d\mathbf{x} \, ds. \end{aligned}$$

Differentiating (2.10) with respect to t we obtain

$$\begin{aligned} &\frac{d}{dt} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} (1 + v^2)^{\sigma/2} f_R(\mathbf{v}, \mathbf{x}, t) \, d\mathbf{v} \, d\mathbf{x} \\ &= \sigma \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} (1 + v^2)^{\sigma/2 - 1} (\mathbf{v} \cdot \mathbf{\Gamma}(\mathbf{x}, t)) f_R(\mathbf{v}, \mathbf{x}, t) \, d\mathbf{v} \, d\mathbf{x} + \frac{d}{dt} I(t). \end{aligned} \tag{2.11}$$

To estimate the last term $dI(t)/dt$, use the following fundamental inequality (Proposition 2.1 in [P2]):

$$\begin{aligned} &(1 + (v')^2)^{\sigma/2} - (1 + v^2)^{\sigma/2} \\ &\leqq K_1 \cdot w \cos \theta (1 + v_*)^{\max(1, \sigma - 1)} (1 + v^2)^{(\sigma-2)/2} - K_2 \cdot w \cos^2 \theta (1 + v^2)^{(\sigma-1)/2}, \end{aligned} \tag{2.12}$$

with positive constants K_1 and K_2 (depending only on σ and κ). Then, as derived in the proof to Theorem 3.1 in [P2], there exist constants $c_0, c_1, c_2 > 0$ such that

$$\text{(2.13)}\quad \frac{d}{dt} I(t) \leqq c_1 \|f_R(\cdot, \cdot, t)\|_{1,\sigma+\gamma-1} + c_2 \|f_R(\cdot, \cdot, t)\|_{1,\sigma-1} - c_0 \|f_R(\cdot, \cdot, t)\|_{1,\sigma+\gamma}.$$

As a consequence of (2.11) and (2.13) we obtain

$$\frac{d}{dt} \|f_R(\cdot, \cdot, t)\|_{1,\sigma} \leqq \sigma_0 \tilde{\Gamma}_0 \|f_R(\cdot, \cdot, t)\|_{1,\sigma-1} + c_1 \|f_R(\cdot, \cdot, t)\|_{1,\sigma+\gamma-1}$$
$$+ c_2 \|f_R(\cdot, \cdot, t)\|_{1,\sigma-1} - c_0 \|f_R(\cdot, \cdot, t)\|_{1,\sigma+\gamma}.$$

Hence,

$$\text{(2.14)}\qquad \frac{d}{dt} \|f_R(\cdot, \cdot, t)\|_{1,\sigma} \leqq c \|f_R(\cdot, \cdot t)\|_{1,\sigma-\delta} - c_0 \|f_R(, \cdot, t)\|_{1,\sigma},$$

where $c > 0$ is a constant and δ satisfies $0 < \delta < \min(\sigma, 1-\gamma)$ and $k\delta = \sigma$ for some nonnegative integer k. It follows from (2.14) that

$$\|f_R(\cdot, \cdot, t)\|_{1,\sigma} \leqq \|\varphi\|_{1,\sigma} + c \int_0^t \exp[-c_0(t-s)] \cdot \|f_R(\cdot, \cdot, s)\|_{1,\sigma-d}\, ds \leqq \text{const} \cdot \|\varphi\|_{1,\sigma},$$

if $\|f_R(\cdot, \cdot, t)\|_{1,\sigma-\delta}$ is globally bounded.

By proceeding step by step, starting with the value $\|f_R(\cdot, \cdot, s)\|_{1,0}$, we obtain consecutively that $\|f_R(\cdot, \cdot, t)\|_{1,n\delta}$, $n = 0, 1, \ldots, k$, are globally bounded on $\mathbb{R}_+$, and that $\|f_R(\cdot, \cdot t)\|_{1,n\delta} \leqq \text{const} \cdot \|\varphi\|_{1,n\delta}$. Finally, let f be the unique mild solution, with initial data φ, of Theorem 2.5. Then by the proof of Theorem 2.5, $f_R(\cdot, \cdot, t) \to f(\cdot, \cdot, t)$ in $L^1_{n\delta}(\mathbb{R}^3 \times \mathbb{R}^3)$ as $R \nearrow \infty$, $t \in \mathbb{R}_+$. This completes the proof. □

3. $L^{1;q} \cap L^{p;q}$-solutions. This section treats existence and global boundedness of solutions to the problem (0.6) in the Banach spaces $L^{1;q}_\sigma(\mathbb{R}^3 \times \mathbb{R}^3) \cap L^{p;q}_\sigma(\mathbb{R}^3 \times \mathbb{R}^3)$, $1 \leqq p, q \leqq \infty$, $\sigma \geqq 0$, where

$$L^{p;q}_\sigma(\mathbb{R}^3 \times \mathbb{R}^3) = \{f : f \text{ measurable on } \mathbb{R}^3 \times \mathbb{R}^3,\ \|f\|_{p,\sigma;q} < \infty\},$$

and

$$\|f(\cdot, \cdot)\|_{p,\sigma;q} = \left(\int_{\mathbb{R}^3} \left(\int_{\mathbb{R}^3} |f(\mathbf{v}, \mathbf{x})(1+v^2)^{\sigma/2}|^q\, d\mathbf{x} \right)^{p/q} d\mathbf{v} \right)^{1/p}.$$

The corresponding positive cones are $L^{p;q}_{\sigma,+} = \{f \in L^{p;q}_\sigma : f \geqq 0 \text{ almost everywhere}\}$. For $\sigma = 0$ we shorten the notation, $L^{p;q} = L^{p;q}_0$ with $\|\cdot\|_{p;q} = \|\cdot\|_{p,0;q}$. We will assume that the distribution function F of the neutral particles, the electric field $\mathbf{\Gamma}$, and the magnetic field $\mathbf{\Omega}$ are spatially homogeneous.

The first part of the section considers (0.6) in $L^{1;q}(\mathbb{R}^3 \times \mathbb{R}^3)$, with $1 \leqq q \leqq \infty$, for a function $B(\theta, w)$ given by (0.2) and (0.3). We prove that a unique solution exists and that its $L^{1;q}$-norm is nonincreasing with respect to time; see Theorems 3.1 and 3.4. Moreover, for higher velocity moments global estimates are obtained for hard interactions; see Theorem 3.3.

The second part gives an L^∞-estimate of the gain operator K, when the function $B(\theta, w)$ satisfies (0.2) and (0.4) (with $\alpha = \gamma$). Together with the results of the first part, this implies global estimates of the solution to (0.6) in $L^{1;q}_+(\mathbb{R}^3 \times \mathbb{R}^3) \cap L^{p;q}(\mathbb{R}^3 \times \mathbb{R}^3)$, where $1 \leqq p, q \leqq \infty$; see Theorems 3.8 and 3.10.

Since $\mathbf{\Gamma} = \mathbf{\Gamma}(t)$, $\mathbf{\Omega} = \mathbf{\Omega}(t)$, the solution of (1.1) is given by

$$\text{(3.1)}\qquad \mathbf{v}(t) = \mathbf{v}(\mathbf{u}, t),\ \mathbf{x}(t) = \mathbf{y} + \int_0^t \mathbf{v}(s)\, ds, \qquad t \in \mathbb{R}_+.$$

For fixed $\mathbf{u}\in\mathbb{R}^3$, it follows by (3.1) that the Jacobian of the transformation $\mathbf{y}\mapsto\mathbf{x}(t)$ equals one for every $t\in\mathbb{R}_+$. Furthermore, by Hypothesis 1.1, the Jacobian of the transformation $\mathbf{u}\mapsto\mathbf{v}(t)$ is also equal to one for every $t\in\mathbb{R}_+$. In the proofs below those transformations will be used without comments.

With F spatially homogeneous, condition (2.1) corresponds to F being measurable on $\mathbb{R}^3\times\mathbb{R}_+$, and

$$(3.2\text{a})\qquad F(\cdot,t)\in L^1_\gamma(\mathbb{R}^3),\quad t\in\mathbb{R}_+\quad \text{if } 0\leqq\gamma\leqq 1,$$

$$(3.2\text{b})\qquad F(\cdot,t)\in L^1(\mathbb{R}^3)\cap L^\infty(\mathbb{R}^3),\quad t\in\mathbb{R}_+\quad \text{if } -3<\gamma<0.$$

Notation. For measurable functions $g:\mathbb{R}^3\times\mathbb{R}^3\to\mathbb{R}$ such that $g(\mathbf{u},\cdot)\in L^q(\mathbb{R}^3)$, almost every $\mathbf{u}\in\mathbb{R}^3$, set

$$g_q(\mathbf{u})=\|g(\mathbf{u},\cdot)\|_q.$$

For measurable functions $g:\mathbb{R}^3\times\mathbb{R}^3\times\mathbb{R}_+\to\mathbb{R}$ such that $g(\mathbf{u},\cdot,t)\in L^q(\mathbb{R}^3)$, almost every $\mathbf{u}\in\mathbb{R}^3$, $t\in\mathbb{R}_+$, set

$$g_q(\mathbf{u},t)=\|g(\mathbf{u},\cdot,t)\|_q,$$

and

$$g_q^{\#}(\mathbf{u},t)=g_q(\mathbf{v}(\mathbf{u},t),t).$$

Consider the L^q-norm, with respect to $\mathbf{y}$, of the gain term $(Kg)^{\#}$. By the Minkowski inequality,

$$\begin{aligned}\|(Kg)^{\#}(\mathbf{u},\cdot,t)\|_q&=\left\|\int_{\mathbb{R}^3}\int_0^{2\pi}\int_0^{\pi/2} g(\mathbf{v}(t)',\mathbf{x}(t),t)F(\mathbf{v}'_*,t)b(\theta)|\mathbf{v}(t)-\mathbf{v}_*|^\gamma\,d\theta\,d\varepsilon\,d\mathbf{v}_*\right\|_q\\&\leqq\int_{\mathbb{R}^3}\int_0^{2\pi}\int_0^{\pi/2}\|g(\mathbf{v}(t)',\mathbf{x}(t),t)\|_q F(\mathbf{v}'_*,t)b(\theta)|\mathbf{v}(t)-\mathbf{v}_*|^\gamma\,d\theta\,d\varepsilon\,d\mathbf{v}_*.\end{aligned}$$

After the transformation $\mathbf{x}(t)\mapsto\mathbf{y}$ we find that $\|g(\mathbf{v}(t)',\mathbf{x}(t),t)\|_q=g_q(\mathbf{v}(t)',t)$. Hence the right-hand side of the inequality above equals $(Kg_q)(\mathbf{v}(t),t)=(Kg_q)^{\#}(\mathbf{u},t)$. We conclude that

$$(3.3)\qquad \|(Kg)^{\#}(\mathbf{u},\cdot,t)\|_q\leqq(Kg_q)^{\#}(\mathbf{u},t)\quad \text{a.e. } \mathbf{u}\in\mathbb{R}^3,\quad t\in\mathbb{R}_+.$$

In the proofs to follow, (3.3) is essential.

In Theorem 3.1 below we prove that the $L^{1;q}$-norm of the mild solution f of problem (0.6) is nonincreasing with time. The main idea of the proof is to take the L^q-norm with respect to $\mathbf{y}$ of the mild equation in the exponential form (1.5). By use of (3.3) it follows that the right-hand side of (1.5) is bounded by the corresponding spatially homogeneous version of (1.5). We then prove that there is a unique spatially homogeneous solution, which is an upper bound for $f_q(\mathbf{u},t)=\|f(\mathbf{u},\cdot t)\|_q$.

THEOREM 3.1. *Let* $-3<\gamma\leqq 1$, $1\leqq q\leqq\infty$, *and suppose* (3.2) *holds. If* $\varphi\in L_+^{1;q}$, *then there exists a mild solution f of problem* (0.6), *with initial data* φ, *such that* $f(\cdot,\cdot,t)\in L_+^{1;q}$, $t\in\mathbb{R}_+$. *Moreover,*

$$\|f(\cdot,\cdot,t)\|_{1;q}\leqq\|\varphi\|_{1;q},\qquad t\in\mathbb{R}_+.$$

Proof. Consider the sequence of iterates given by (2.3). If, in the initial function, φ is replaced by φ_n, where $\varphi_n(\mathbf{u},\mathbf{y})=\varphi(\mathbf{u},\mathbf{y})$ if $|\mathbf{u}|\leqq n$, and $\varphi_n(\mathbf{u},\mathbf{y})=0$ otherwise, we still obtain a nonnegative and nondecreasing sequence $\{f_n\}_0^\infty$. By (3.3) and the Minkowski inequality it follows that the sequence satisfies the inequality

$$\begin{aligned}(3.4)\quad f^{\#}_{n+1,q}(\mathbf{u},t)&\leqq \varphi_{n,q}(\mathbf{u})\exp\left(-\int_0^t \nu^{\#}(\mathbf{u},\tau)\,d\tau\right)\\ &\quad+\int_0^t (Kf_{n,q})^{\#}(\mathbf{u},s)\exp\left(-\int_s^t \nu^{\#}(\mathbf{u},\tau)\,d\tau\right)ds,\quad \mathbf{u}\in\mathbb{R}^3,\quad t\in\mathbb{R}_+;\end{aligned}$$

here and below $\varphi_{n,q}=(\varphi_n)_q$, and $f_{n,q}=(f_n)_q$. Now, since $\nu\geqq 0$, we obtain by induction that

$$(3.5)\qquad f^{\#}_{n,q}(\cdot,t)\in L^1_\sigma(\mathbb{R}^3)\quad\text{for all } \sigma\geqq 0,\quad t\in\mathbb{R}_+.$$

Next, consider the spatially homogeneous sequence $\{g_{n,m}\}_{m=0}^\infty$ (depending on q) defined by

$$\begin{aligned}g^{\#}_{n,0}(\mathbf{u},t)&=f^{\#}_{n,q}(\mathbf{u},t),\\ (3.6)\quad g^{\#}_{n,m+1}(\mathbf{u},t)&=\varphi_{n,q}(\mathbf{u})\exp\left(-\int_0^t \nu^{\#}(\mathbf{u},\tau)\,d\tau\right)\\ &\quad+\int_0^t (Kg_{n,m})^{\#}(\mathbf{u},s)\exp\left(-\int_s^t \nu^{\#}(\mathbf{u},\tau)\,d\tau\right)ds,\quad \mathbf{u}\in\mathbb{R}^3,\quad t\in\mathbb{R}_+.\end{aligned}$$

Notice that (3.4) implies that $\{g_{n,m}\}$ is nondecreasing, and (3.5) implies that $g_{n,m}(\cdot,t)\in L^1_\sigma(\mathbb{R}^3)$, $\sigma\geqq 0$, $t\in\mathbb{R}_+$. We now conclude by the proof of Theorem 2.3 that $g_{n,m}$ converges when $m\to\infty$ to a nonnegative mild solution g_n of the spatially homogeneous version of (0.6) with initial data $\varphi_{n,q}$; moreover, $g_n(\cdot,t)\in L^1_+(\mathbb{R}^3)$, $t\in\mathbb{R}_+$, and

$$(3.7)\qquad \|g_n(\cdot,t)\|_1\leqq\|\varphi_{n,q}\|_1\leqq\|\varphi_q\|_1,\qquad t\in\mathbb{R}_+.$$

By the construction of g_n we have $f_{n,q}=g_{n,0}\leqq g_n$. Hence

$$\|f_n(\cdot,\cdot,t)\|_{1;q}=\|f_{n,q}(\cdot,t)\|_1\leqq\|g_n(\cdot,t)\|_1\leqq\varphi_q\|_1=\|\varphi\|_{1;q},\qquad t\in\mathbb{R}_+.$$

Finally, set $f(\mathbf{v},\mathbf{x},t)=\lim_{n\to\infty}f_n(\mathbf{v},\mathbf{x},t)$, $(\mathbf{v},\mathbf{x})\in\mathbb{R}^3\times\mathbb{R}^3$, $t\in\mathbb{R}_+$. Then f is measurable and by monotone convergence

$$\|f(\cdot,\cdot,t)\|_{1;q}=\lim_{n\to\infty}\|f_n(\cdot,\cdot,t)\|_{1;q}\leqq\|\varphi\|_{1;q},\qquad t\in\mathbb{R}_+.$$

Moreover, letting $n\to\infty$ in (2.4), we establish that f is a mild solution of problem (0.6). □

Remark. (i) For later reference let $g(\mathbf{v},t)=\lim_{n\to\infty}g_n(\mathbf{v},t)$, $\mathbf{v}\in\mathbb{R}^3$, $t\in\mathbb{R}_+$. Then g is a mild solution to (0.6) with initial data φ_q, and if $\text{ess-sup}_{t\in\mathbb{R}_+}\|F(\cdot,t)\|_{1,2+\max(-1,\gamma)}<\infty$, $-3<\gamma<1$, then by Theorem 2.5 $\|g(\cdot,t)\|_1=\|\varphi_q\|_1$, $t\in\mathbb{R}_+$. Moreover, by the uniqueness part of Theorem 2.5, it follows that, with $g_0=0$ instead of $g_{n,0}=f_{n,q}$ and φ_q instead of $\varphi_{n,q}$ in (3.6), we obtain a sequence $\{g_m\}$ which converges to the same solution g.

(ii) In the theorems that follow the mild solution of (0.6) is the (solution that here was constructed as the) pointwise limit of the iterate functions $\{f_n\}$ defined by (2.3) (with φ replaced by φ_n).

Next, the solution of Theorem 3.1 is studied for higher velocity moments. Local boundedness in the case $-3<\gamma<1$ and global boundedness in the case $0\leqq\gamma\leqq 1$ follow immediately from Theorems 2.4 and 2.6, respectively.

THEOREM 3.2. *Let* $-3<\gamma<1$, $\sigma_0\geqq 2$. *When* $-1\leqq\gamma$, *let* (3.2) *hold together with* $\text{ess-sup}_{t\in\mathbb{R}_+}\|F(\cdot,t)\|_{1,\sigma_0+\gamma}<\infty$; *otherwise, let* $\text{ess-sup}_{t\in\mathbb{R}_+}\|F(\cdot,t)\|_{1\cap\infty,\sigma_0-1}<\infty$. *Let* $\varphi\in L^{1;q}_{\sigma,+}$ *with* $1\leqq q\leqq\infty$, *and* $0\leqq\sigma\leqq\sigma_0$. *Then the mild solution f given in Theorem* 3.1 *satisfies* $f(\cdot,\cdot,t)\in L^{1;q}_{\sigma}$, $t\in\mathbb{R}_+$, *and*

$$\|f(\cdot,\cdot,t)\|_{1,\sigma;q}\leqq\|\varphi\|_{1,\sigma;q}\cdot\exp(\psi(t)),\qquad t\in\mathbb{R}_+,$$

where $\psi(t)=\sigma_0(\int_0^t\tilde{\Gamma}(s)\,ds+ct)$ *and c is a positive constant.*

THEOREM 3.3. *In addition to the assumptions of Theorem* 3.2 *suppose that* $0\leqq\gamma<1$, $\tilde{\Gamma}_0<\infty$, *and that* $\text{ess-inf}_{t\in\mathbb{R}_+}\|F(\cdot,t)\|_1>0$. *Then the mild solution f satisfies*

$$\|f(\cdot,\cdot,t)\|_{1,\sigma;q}\leqq c\cdot\|\varphi\|_{1,\sigma;q},\quad t\in\mathbb{R}_+,$$

where c is a positive constant independent of φ.

Proofs of Theorems 3.2 *and* 3.3. Let f be the mild solution given in Theorem 3.1. Let g be the space homogeneous mild solution with initial data φ_q. Then by the proof of Theorem 3.1,

$$\|f(\mathbf{v},\cdot,t)\|_q=f_q(\mathbf{v},t)\leqq g(\mathbf{v},t)\quad\text{a.e. }\mathbf{v}\in\mathbb{R}^3,\quad t\in\mathbb{R}_+. \tag{3.8}$$

Finally, (3.8), together with Theorems 2.4 and 2.6, implies Theorems 3.2 and 3.3, respectively. □

The solution f in Theorems 3.1–3.3 is unique in the following sense.

THEOREM 3.4. *Suppose that* $-3<\gamma<1$, (3.2) *holds*, $\text{ess-sup}_{t\in\mathbb{R}+}\|F(\cdot,t)\|_{1,2+\gamma}<\infty$ *when* $0<\gamma<1$, *and that* $\varphi\in L^{1;q}_{\max(0,\gamma),+}$. *If f is the mild solution of* (0.6) *obtained in Theorem* 3.1 *and if* $\tilde{f}$ *is any mild solution to* (0.6) *such that* $\tilde{f}(\cdot,\cdot,t)\in L^{1;q}_{+}$ *and* $\nu(\cdot,t)\tilde{f}_q(\cdot,t)\in L^1$, *then*

$$\tilde{f}(\mathbf{u},\mathbf{y},t)=f(\mathbf{u},\mathbf{y},t)\quad\textit{for a.e. }(\mathbf{u},\mathbf{y})\in\mathbb{R}^3\times\mathbb{R}^3,\qquad t\in\mathbb{R}_+.$$

Proof. By differentiating the $L^{1;q}$-norm (with respect to time) of the difference between f and $\tilde{f}$ we obtain (cf. [Gu2, Lemma 1.4])

$$\begin{aligned}\frac{d}{dt}\|f^*(t)-\tilde{f}^*(t)\|_{1;q}=&\int_{\mathbb{R}^3}(((f-\tilde{f})^*)_q(\mathbf{u},t))^{1-q}\\&\cdot\int_{\mathbb{R}^3}\operatorname{sgn}(f-\tilde{f})^*(\mathbf{u},\mathbf{y},t)\\&\cdot(|f-\tilde{f}|^*(\mathbf{u},\mathbf{y},t))^{q-1}\frac{d}{dt}(f-\tilde{f})^*(\mathbf{u},\mathbf{y},t)\,d\mathbf{y}\,d\mathbf{u}.\end{aligned}$$

We have

$$\frac{d}{dt}(f-\tilde{f})^*(\mathbf{u},\mathbf{y},t)=(K(f-\tilde{f}))^*(\mathbf{u},\mathbf{y},t)-\nu^*(\mathbf{u},t)(f-\tilde{f})^*(\mathbf{u},\mathbf{y},t),$$

$$\text{for a.e. }(\mathbf{u},\mathbf{y})\in\mathbb{R}^3\times\mathbb{R}^3,\qquad t\in\mathbb{R}_+.$$

Hence

$$\begin{aligned}&\frac{d}{dt}\|f^*(t)-\tilde{f}^*(t)\|_{1;q}\\&\leqq\int_{\mathbb{R}^3}((f-\tilde{f})^*_q(\mathbf{u},t))^{1-q}\int_{\mathbb{R}^3}(|f-\tilde{f}|^*(\mathbf{u},\mathbf{y},t))^{q-1}(K|f-\tilde{f}|)^*(\mathbf{u},\mathbf{y},t)\,d\mathbf{y}\,d\mathbf{u}\\&\quad-\int_{\mathbb{R}^3}\nu^*(\mathbf{u},t)((f-\tilde{f})^*_q(\mathbf{u},t))^{1-q}\int_{\mathbb{R}^3}(|f-\tilde{f}|^*(\mathbf{u},\mathbf{y},t))^q\,d\mathbf{y}\,d\mathbf{u}.\end{aligned}$$

(Observe here and also below that by the transformation $\mathbf{y}\to\mathbf{x}(t)$ we get $((f-\tilde{f})^{\#})_q(\mathbf{u}, t)=(f-\tilde{f})^{\#}_q(\mathbf{u}, t)$.)

Now, using Hölder's inequality on the first term on the right-hand side and observing that the second term equals $\int_{\mathbb{R}^3} \nu^{\#}(\mathbf{u}, t)(f-\tilde{f})^{\#}_q(\mathbf{u}, t)\, d\mathbf{u}$ it follows that

$$\frac{d}{dt}\|f^{\#}(t)-\tilde{f}^{\#}(t)\|_{1;q} \leqq \int_{\mathbb{R}^3} \|(K(f-\tilde{f}))^{\#}\|_q(\mathbf{u}, t)\, d\mathbf{u} - \int_{\mathbb{R}^3} \nu^{\#}(\mathbf{u}, t)(f-\tilde{f})^{\#}_q(\mathbf{u}, t)\, d\mathbf{u}.$$

Finally, by (3) the first term on the right is less than or equal to $\int_{\mathbb{R}^3} (K(f-\tilde{f})_q)^{\#}(\mathbf{u}, t)$; so after the transformation $\mathbf{u}\to\mathbf{v}(t)$ it follows by (2.2) that the right-hand side is not greater than zero. Thus the theorem is proved. Note also that by Theorem 3.2 all integrals involved exist. □

For the next part we need some *L^∞-properties* of the gain operator K. Here the dependence on $\mathbf{x}$ and t plays no role and will, therefore, be left out.

The following transformation of Kg is used. Let $E_{\mathbf{v},\bar{\mathbf{v}}}$ denote the plane in $\mathbb{R}^3$ through $\tilde{\mathbf{v}}=((1+\kappa)/2)\mathbf{v}+((1-\kappa)/2)\bar{\mathbf{v}}$ and orthogonal to $\mathbf{v}-\bar{\mathbf{v}}$;

$$E_{\mathbf{v},\bar{\mathbf{v}}}=\{\mathbf{u}\in\mathbb{R}^3;\ (\mathbf{u}-\tilde{\mathbf{v}})\cdot(\mathbf{v}-\bar{\mathbf{v}})=0\}.$$

The corresponding Lebesgue measure is $\int dE$.

LEMMA 3.5. *The gain term can be represented as follows*:

$$Kg(\mathbf{v})=\left(\frac{1+\kappa}{2}\right)^{1+\gamma}\int_{\mathbb{R}^3} g(\mathbf{v}')|\mathbf{v}-\mathbf{v}'|^{\gamma-2}\int_{E_{\mathbf{v},\mathbf{v}'}} F(\mathbf{v}'_*)b(\theta)(\sin\theta)^{-1}(\cos\theta)^{-\gamma}\, dE'_*\, d\mathbf{v}'.$$

Proof. The proof is a straightforward generalization of the proof in [Ca]. □

Below we consider $B(\theta, w)$ satisfying (0.2) and (0.4) (with $\alpha=\gamma$). Let $h_{\sigma,R}$ be as in (2.6), and denote $\|g\|_{\infty,\sigma,R}=\|h_{\sigma,R}g\|_\infty$ for measurable functions $g:\mathbb{R}^3\to\mathbb{R}$.

LEMMA 3.6. *Suppose* $-1<\gamma\leqq 1$, $F\in L^\infty_{2+\sigma_1}(\mathbb{R}^3)$, *where* $\sigma_1>0$, $g\in L^1_{\sigma,+}(\mathbb{R}^3)\cap L^\infty(\mathbb{R}^3)$, *where* $0\leqq\sigma\leqq\sigma_1$. *Then, for every* $\varepsilon>0$ *there is a constant* c_ε *such that for* $1<R<\infty$,

$$Kg(\mathbf{v})h_{\sigma,R}(v)\leqq c_\varepsilon\|g\|_{1,\sigma}+\varepsilon\|g\|_{\infty,\sigma,R},\qquad \mathbf{v}\in\mathbb{R}^3. \tag{3.9}$$

In particular, if $g\in L^1_+(\mathbb{R}^3)\cap L^\infty(\mathbb{R}^3)$, *then it holds that*

$$Kg(\mathbf{v})\leqq c_\varepsilon\|g\|_1+\varepsilon\|g\|_\infty,\qquad \mathbf{v}\in\mathbb{R}^3. \tag{3.10}$$

The constant c_ε *may be chosen such that it depends only on* ε, R, σ, σ_1, *and the* $L^\infty_{\sigma_1}$*-norm of* F.

Proof. Cf. [Gu1], or see [CGP]. □

In the next lemma we present (in the hard interaction case) a lower bound of the collision frequency

$$\nu(\mathbf{v}, t)=2\pi\int_0^{\pi/2} b(\theta)\, d\theta\int_{\mathbb{R}^3} F(\mathbf{v}_*, t)|\mathbf{v}-\mathbf{v}_*|^\gamma\, d\mathbf{v}_*.$$

LEMMA 3.7. *Suppose* $\gamma\geqq 0$, *and that* (3.2) *holds*, ess-sup$_{t\in\mathbb{R}_+}$ $\|F(\cdot, t)\|_\infty<\infty$, *and* $\inf_{t\in\mathbb{R}_+}\|F(\cdot, t)\|_1>0$. *Then*

$$\nu(\mathbf{v}, t)\geqq c_1,\quad \mathbf{v}\in\mathbb{R}^3,\quad t\in\mathbb{R}_+,\quad \textit{for some positive constant } c_1. \tag{3.11}$$

Proof. Cf. [Gu1, Lemma 3.1]. □

Our next theorem concerns existence and global estimates of a solution to (0.6) in the Banach spaces $L^{1;q}_+\cap L^{p;q}$, $1\leqq p\leqq\infty$. The $L^{1;q}\cap L^{\infty;q}$ case is obtained by exploiting the L^∞-estimate of the gain operator K given in Lemma 3.6. Combining this with the known $L^{1;q}$ case in Theorem 3.1, the final result follows by linear interpolation.

THEOREM 3.8. *Let* $0\leqq\gamma\leqq 1$, $1\leqq p, q\leqq\infty$. *Suppose* (3.2) *holds together with* ess-sup$_{t\in\mathbb{R}_+}\|F(\cdot,t)\|_{\infty,\sigma}<\infty$ *for some* $\sigma>2$, *and* ess-inf$_{t\in\mathbb{R}_+}\|F(\cdot,t)\|_1>0$. *If* $\varphi\in L_+^{1;q}\cap L^{p;q}$, *then there exists a mild solution f of the problem* (0.6) *with initial data* φ, *such that* $f(\cdot,\cdot,t)\in L_+^{1;q}\cap L^{p;q}$, $t\in\mathbb{R}_+$,

$$\|f(\cdot,\cdot,t)\|_{1;q}\leqq|\varphi\|_{1;q},\qquad t\in\mathbb{R}_+,$$

and

$$\|f(\cdot,\cdot,t)\|_{p;q}\leqq C(\|\varphi\|_{1;q}+\|\varphi\|_{p;q}),\qquad t\in\mathbb{R}_+.$$

Here C is a constant not depending on φ.

Proof. We first prove the theorem in the case $p=\infty$. By Theorem 3.1 there exists a mild solution $f(\cdot,\cdot,t)\in L_+^{1;q}$, $t\in\mathbb{R}_+$, to (0.6a) with initial data φ. Let g be the spatially homogeneous mild solution to (0.6a) with initial data φ_q. Then by construction of f and g (cf. the remark after Theorem 3.1)

$$f_q(\mathbf{v},t)\leqq g(\mathbf{v},t)\quad \text{a.e. } \mathbf{v}\in\mathbb{R}^3,\qquad t\in\mathbb{R}_+. \tag{3.12}$$

Consider the sequence $\{g_m\}_{m=0}^{\infty}$ defined by

$$\begin{aligned} &g_0^{\#}(\mathbf{u},t)=0,\\ &g_{m+1}^{\#}(\mathbf{u},t)=\varphi_q(\mathbf{u})\exp\left(-\int_0^t \nu^{\#}(\mathbf{u},\tau)\,d\tau\right)\\ &\qquad+\int_0^t (Kg_m)^{\#}(\mathbf{u},s)\exp\left(-\int_s^t \nu^{\#}(\mathbf{u},\tau)\,d\tau\right)ds,\\ &\qquad(\mathbf{u},t)\in\mathbb{R}^3\times\mathbb{R}_+. \end{aligned}$$

Applying the upper bound (3.10) on Kg_m and the lower bound (3.11) of ν, we get

$$g_{m+1}^{\#}(\mathbf{u},t)\leqq\varphi_q(\mathbf{u})+\frac{1}{c_1}\left(c_\varepsilon\|\varphi_q\|_1+\varepsilon\ \underset{t\in\mathbb{R}_+}{\text{ess-sup}}\ \|g_m(\cdot,t)\|_\infty\right). \tag{3.13}$$

Since $g_0=0$, it follows by induction that ess-sup$_{t\in\mathbb{R}_+}\|g_m(\cdot,t)\|_\infty<\infty$ for $m=0,1,2,\dots$. Now with $\varepsilon=c_1/2$, after a rearrangement of (3.13), using $g_m\leqq g_{m+1}$,

$$\|g_m(\cdot,t)\|_\infty\leqq 2\|\varphi_q\|_\infty+\left(\frac{2c_\varepsilon}{c_1}\right)\|\varphi_q\|_1,\qquad m=0,1,2,\dots,\quad t\in\mathbb{R}_+.$$

The right-hand side is independent of m and g_m increases pointwise towards g (with the exception of a null set) when $m\to\infty$. Hence

$$\|g(\cdot,t)\|_\infty\leqq 2\|\varphi_q\|_\infty+\left(\frac{2c_\varepsilon}{c_1}\right)\|\varphi_q\|_1,\qquad t\in\mathbb{R}_+. \tag{3.14}$$

Moreover, by Theorem 2.5,

$$\|g(\cdot,t)\|_1=\|\varphi_q\|_1,\qquad t\in\mathbb{R}_+. \tag{3.15}$$

Next, for any fixed $t\in\mathbb{R}_+$, let T_t be the linear operator such that, given $\varphi\in L_+^1(\mathbb{R}^3)$, $T_t\varphi$ is the spatially homogeneous mild solution at time t of (0.6a) with initial data φ. Then, by (3.14) and (3.15) T_t is a bounded linear operator on $L_+^1(\mathbb{R}^3)$ and on $L_+^1(\mathbb{R}^3)\cap L^\infty(\mathbb{R}^3)$. Using the extended interpolation theorem of Riesz-Thorin (cf. [BL, Thm. 3.1.2]), the observation that the K-functional (cf. [BL]) is not altered when considering only nonnegative functions, and that the interpolation space $(L^1(\mathbb{R}^3), L^1(\mathbb{R}^3)\cap L^\infty(\mathbb{R}^3))_{1/p',p}=L^1(\mathbb{R}^3)\cap L^p(\mathbb{R}^3)$, where $1/p+1/p'=1$ (cf. [Gul]), we obtain that T_t is bounded on $L^1_+(\mathbb{R}^3) \cap L^p(\mathbb{R}^3)$ for any p with $1 \leqq p \leqq \infty$. Furthermore, for some constant C,

$$\|T_t\varphi\|_p \leqq C(\|\varphi\|_1 + \|\varphi\|_p), \qquad t \in \mathbb{R}_+. \tag{3.16}$$

From here the theorem follows by (3.12), (3.15), and (3.16) with $\varphi = \varphi_q$. □

The last theorem of this section examines higher moments of the solution of Theorem 3.8. In the proof we employ the following elementary inequality of Gronwall type; cf. [Gu1].

LEMMA 3.9. *Let h_1, h_2, and g be nonnegative and continuous functions on $\mathbb{R}_+$. If h_1 is positive, g is locally absolutely continuous on $\mathbb{R}_+$, and*

$$\frac{dg}{dt} + h_1 g \leqq h_2 \quad \textit{for a.e. } t \in \mathbb{R}_+,$$

then

$$\sup_{t\in\mathbb{R}_+} g(t) \leqq g(0) + \sup_{t\in\mathbb{R}_+}\left(\frac{h_2(t)}{h_1(t)}\right).$$

THEOREM 3.10. *Let $0 \leqq \gamma < 1$, $0 \leqq \bar\sigma \leqq \sigma$, $1 \leqq p, q \leqq \infty$. Suppose* (3.2) *holds together with* $\text{ess-sup}_{t\in\mathbb{R}_+} \|F(\cdot, t)\|_{\infty, 2+\max(\sigma,\gamma)} < \infty$, *and* $\text{ess-inf}_{t\in\mathbb{R}_+} \|F(\cdot, t)\|_1 > 0$. *If $\varphi \in L^{1,q}_{\sigma,+} \cap L^{p;q}_{\bar\sigma}$, then there exists a mild solution f of the problem* (0.6) *such that $f(\cdot, \cdot, t) \in L^{1;q}_{\sigma,+} \cap L^{p;q}_{\bar\sigma}$, $t \in \mathbb{R}_+$,*

$$\|f(\cdot,\cdot,t)\|_{1,\bar\sigma;q} \leqq C_0\|\varphi\|_{1,\bar\sigma;q}, \qquad t \in \mathbb{R}_+,$$

$$\|f(\cdot,\cdot,t)\|_{p,\bar\sigma,q} \leqq C_1(\|\varphi\|_{1,\bar\sigma;q} + \|\varphi\|_{p,\bar\sigma;q}), \qquad t \in \mathbb{R}_+.$$

Here C_0 and C_1 are constants not depending on φ.

Proof. As in the proof of Theorem 3.8 we first prove the theorem in the case $p = \infty$. Let f and g be as in the proof of Theorem 3.8. Then (cf. Theorem 3.3) there holds

$$\|f(\cdot,\cdot,t)\|_{1,\sigma;q} \leqq \|g_i(\cdot,t)\|_{1,\sigma} \leqq C \cdot \|\varphi\|_{1,\sigma;q}, \qquad t \in \mathbb{R}_+ \tag{3.17}$$

for some constant C not dependent on φ. Moreover, since g is a spatially homogeneous mild solution of (0.6a) with initial data φ_q, it follows that for almost every $\mathbf{u} \in \mathbb{R}^3$, $g^\#(\mathbf{u}, \cdot)$ is locally absolutely continuous on $\mathbb{R}_+$. Hence, for any σ: $0 \leqq \sigma \leqq \bar\sigma$,

$$\begin{aligned}\frac{\partial}{\partial t}[g^\#(\mathbf{u},t)h^\#_{\sigma,R}(\mathbf{u},t)] + \nu^\#(\mathbf{u},t)g^\#(\mathbf{u},t)h^\#_{\sigma,R}(\mathbf{u},t)\\ = (Kg)^\#(\mathbf{u},t)h^\#_{\sigma,R}(\mathbf{u},t) + g^\#(\mathbf{u},t)\frac{\partial}{\partial t}h^\#_{\sigma,R}(\mathbf{u},t)\end{aligned} \tag{3.18}$$

for $\mathbf{u} \in \mathbb{R}^3 \backslash M$, $t \in \mathbb{R}_+ \backslash M_{\mathbf{u}}$, with $\operatorname{mes} M = \operatorname{mes} M_{\mathbf{u}} = 0$. Next by (2.7), (3.9), (3.11), (3.17), and (3.18) we obtain

$$\begin{aligned}\frac{\partial}{\partial t}[g^\#(\mathbf{u},t)h^\#_{\sigma,R}(\mathbf{u},t)] + c_1 g^\#(\mathbf{u},t)h^\#_{\sigma,R}(\mathbf{u},t)\\ \leqq c_\varepsilon\|\tilde\varphi_q\|_{1,\tilde\varphi} + \varepsilon\|g(\cdot,t)\|_{\infty,\sigma,R} + \sigma\tilde\Gamma_0\|g(\cdot,t)\|_{\infty,\max(\sigma-1,0)}.\end{aligned} \tag{3.19}$$

Thus, by Theorem 3.8, the right-hand side of (3.19) with $0 \leqq \sigma \leqq \min(1, \bar\sigma)$ is finite. Applying Lemma 3.9 we find after a rearrangement with $\varepsilon = c_1/2$ and in the limit $R \to \infty$,

$$\sup_{t\in\mathbb{R}_+}\|g(\cdot,t)\|_{\infty,\sigma} \leqq 2\|\varphi_q\|_{\infty,\bar\sigma} + \frac{2}{c_1}\left(c_\varepsilon\|\varphi_q\|_{1,\bar\sigma} + \bar\sigma\tilde\Gamma_0\sup_{t\in\mathbb{R}_+}\|g(\cdot,t)\|_{\infty,\max(\sigma-1,0)}\right). \tag{3.20}$$

Now, (3.20) shows that the right-hand side of (3.19) is finite for $0 \leqq \sigma \leqq \min(2, \bar\sigma)$. Hence, repeating the above it follows, by induction, that (3.20) is valid for $0 \leqq \sigma \leqq \bar\sigma$.

Thus, using Theorem 3.8 and (3.17), there is a constant $\tilde{C}$ not dependent on φ such that

$$\|g(\cdot,t)\|_{\infty,\tilde\sigma} \leqq \tilde{C}(\|\varphi_q\|_{1,\tilde\sigma}+\|\varphi_q\|_{\infty,\tilde\sigma}), \qquad t\in\mathbb{R}_+. \tag{3.21}$$

Finally, by (3.17) and (3.21), the linear operator T_t defined at the end of the proof of Theorem 3.8 is bounded on $L^1_{\tilde\sigma,+}(\mathbb{R}^3)$ and on $L^\infty_{\tilde\sigma,+}(\mathbb{R}^3)$. Thus $g(\cdot,t)\in L^1_{\tilde\sigma,+}(\mathbb{R}^3)\cap L^p_{\tilde\sigma}(\mathbb{R}^3)$, $t\in\mathbb{R}_+$, and there is a constant C not dependent on φ such that

$$\|g(\cdot,t)\|_{p,\tilde\sigma} \leqq C(\|\varphi_q\|_{1,\tilde\sigma}+\|\varphi_q\|_{p,\tilde\sigma}), \qquad t\in\mathbb{R}_+. \tag{3.22}$$

By (3.12) and (3.22) the theorem is proved. □

4. L^p-solutions in the soft interaction case. The main result of this section is Theorem 4.6, concerning existence and uniqueness of solutions in $L^p(\mathbb{R}^3\times\mathbb{R}^3)$, $1\leqq p\leqq\infty$. Here Lemma 4.3 plays an important role implying that the collision operator is bounded on L^p if the ratio of masses (κ) does not equal one (see Proposition 4.4). A similar approach was used in [Mo] to treat L^1 and L^∞ cases under some hypotheses more restrictive than ours.

LEMMA 4.1. *Let h be a measurable function on $\mathbb{R}^N\times\mathbb{R}^N$ and H the linear integral operator with the kernel h (i.e., $(Hg)(\mathbf{v})=\int_{\mathbb{R}^N}h(\mathbf{v},\mathbf{u})g(\mathbf{u})\,d\mathbf{u}$). If* ess-sup$_{\mathbf{u}\in\mathbb{R}^N}\int_{\mathbb{R}^N}|h(\mathbf{v},\mathbf{u})|\,d\mathbf{v}<\infty$, *then H is a bounded mapping from $L^1(\mathbb{R}^N)$ into itself. If in addition* ess-sup$_{\mathbf{v}\in\mathbb{R}^N}\int_{\mathbb{R}^N}|h(\mathbf{v},\mathbf{u})|\,d\mathbf{v}<\infty$, *then for every $1\leqq p\leqq\infty$, H is a bounded mapping from $L^p(\mathbb{R}^N)$ into itself.*

Proof. See, e.g., [Ed]. □

It is well known that the gain collision operator K has the following form (cf. [Ce]),

$$(Kg)(\mathbf{v},\mathbf{x},t)=\int_{\mathbb{R}^3}k(\mathbf{v},\mathbf{v}')g(\mathbf{v}',\mathbf{x},t)\,d\mathbf{v}'.$$

If the distribution function for the neutral particles F depends also on $\mathbf{x}$ and/or t, so does the kernel k; since F is nonnegative, the same holds for k. Recall that (cf. [Ce])

$$\int_{\mathbb{R}^3}k(\mathbf{v}',\mathbf{v})\,d\mathbf{v}'=\nu(\mathbf{v}), \qquad \mathbf{v}\in\mathbb{R}^3. \tag{4.1}$$

Throughout this section it is tacitly assumed that B satisfies (0.2) and (0.4) (with $\alpha=\gamma$).

LEMMA 4.2. *Suppose that $-1<\gamma\leqq 1$ and (2.1) holds. If also $\kappa\neq 1$, then*

$$\int_{\mathbb{R}^3}k(\mathbf{v},\mathbf{v}')\,d\mathbf{v}'\leqq b_0\left|\frac{1+\kappa}{1-\kappa}\right|^{1+\gamma}\int_{\mathbb{R}^3}\bar{k}(\mathbf{v}',\mathbf{v})\,d\mathbf{v}'.$$

Here $\bar{k}$ is the kernel of the gain collision operator with $b(\theta)=\sin\theta(\cos\theta)^\gamma$.

Proof. According to (0.4) and Lemma 3.5

$$\int_{\mathbb{R}^3}k(\mathbf{v},\mathbf{v}')\,d\mathbf{v}'\leqq b_0\left(\frac{1+\kappa}{2}\right)^{1+\gamma}\int_{\mathbb{R}^3}|\mathbf{v}-\mathbf{v}'|^{\gamma-2}\int_{\mathbf{u}\cdot(\mathbf{v}-\mathbf{v}')=0}F(\tilde{\mathbf{v}}+\mathbf{u})\,dE_{\mathbf{u}}\,d\mathbf{v}',$$

where $\tilde{\mathbf{v}}=((1+\kappa)/2)\mathbf{v}+((1-\kappa)/2)\mathbf{v}'$. Set $\mathbf{r}=\mathbf{v}'-\mathbf{v}$, and use $\mathbf{r}$ as the variable integration instead of $\mathbf{v}'$. Introduce spherical coordinates $(r,\boldsymbol{\alpha})$ for $\mathbf{r}$. Then, by Fubini's theorem,

$$\int_{\mathbb{R}^3}k(\mathbf{v},\mathbf{v}')\,d\mathbf{v}'\leqq b_0\left(\frac{1+\kappa}{2}\right)^{1+\gamma}\int_S\int_{\mathbf{u}\cdot\boldsymbol{\alpha}=0}\int_0^\infty r^\gamma\cdot F\left(\mathbf{v}+\frac{1-\kappa}{2}r\boldsymbol{\alpha}+\mathbf{u}\right)dr\,dE_{\mathbf{u}}\,d\boldsymbol{\alpha},$$

where S denotes the unit sphere in $\mathbb{R}^3$.

Substituting $|(1-\kappa)/(1+\kappa)|r\mapsto r$, the right-hand side equals

$$b_0\left(\frac{1+\kappa}{2}\right)^{1+\gamma}\left|\frac{1+\kappa}{1-\kappa}\right|^{1+\gamma}\int_S\int_{\mathbf{u}\cdot\boldsymbol{\alpha}=0}\int_0^\infty r^\gamma F\left(\mathbf{v}+\frac{1+\kappa}{2}r\boldsymbol{\alpha}+\mathbf{u}\right)dr\,dE_{\mathbf{u}}\,d\boldsymbol{\alpha}.$$

Finally, noticing that $\mathbf{v}+((1+\kappa)/2)\mathbf{r}=((1+\kappa)/2)\mathbf{v}'+((1-\kappa)/2)\mathbf{v}$, the proof is completed. □

LEMMA 4.3. *Suppose that* $-1<\gamma\leqq 1$. *If* $F\in L^{\infty}_{\sigma_0}(\mathbb{R}^3)$ *for some* $\sigma_0>2$, *then*

$$\mathbf{k}(\mathbf{v},\mathbf{v}')\leqq b_0\|F\|_{\infty,\sigma_0}\left(\frac{1+\kappa}{2}\right)^{1+\gamma}\frac{2\pi}{\sigma_0-2}|\mathbf{v}-\mathbf{v}'|^{\gamma-2}\left(1+\frac{1}{4}\left(\kappa|\mathbf{v}-\mathbf{v}'|+\frac{v^2-v'^2}{|\mathbf{v}-\mathbf{v}'|}\right)^2\right)^{-(\sigma_0-2)/2},$$

$\mathbf{v},\mathbf{v}'\in\mathbb{R}^3$, $\mathbf{v}\neq\mathbf{v}'$.

Proof. Cf. [Ch, Lemma 4.4]. □

PROPOSITION 4.4. *Suppose that* $-1<\gamma\leqq 0$, $F\in L^{\infty}_{\sigma_0}$ *with* $\sigma_0>3$, *and* $0\leqq\sigma<\sigma_0-3-\gamma$. *Then K is a bounded mapping from* $L^1_{\sigma}(\mathbb{R}^3)$ *into itself. If moreover* $\kappa\neq 1$, *then K is a bounded mapping from* $L^p_{\sigma}(\mathbb{R}^3)$ *into itself for every p*: $1\leqq p\leqq\infty$.

Proof. To prove the latter assertion of the proposition, it is sufficient to show that the linear integral operator $\tilde{K}$ with the kernel

$$\tilde{k}(\mathbf{v},\mathbf{v}')=\left(\frac{1+v^2}{1+v'^2}\right)^{\sigma/2}k(\mathbf{v},\mathbf{v}'),\qquad \mathbf{v},\mathbf{v}'\in\mathbb{R}^3,\quad \mathbf{v}\neq\mathbf{v}'$$

represents a bounded mapping from $L^p(\mathbb{R}^3)$ into itself for every $1\leqq p\leqq\infty$. Denoting $\mathbf{r}=\mathbf{v}'-\mathbf{v}$ again in view of

$$\frac{1+v^2}{1+v'^2}\leqq\frac{2+(v'+r)^2}{1+v'^2}\leqq 2\frac{1+v'^2+r^2}{1+v'^2}\leqq 2(1+r^2),\qquad \mathbf{v},\mathbf{v}'\in\mathbb{R}^3,$$

we find, using Lemma 4.3, that

$$\tilde{k}(\mathbf{v},\mathbf{v}')\leqq\frac{c}{r^{2-\gamma}}(1+r^2)^{1-\sigma_0/2+\sigma/2},\qquad \mathbf{v},\mathbf{v}'\in\mathbb{R}^3,\quad v\geqq v'. \tag{4.2}$$

On the other hand,

$$\tilde{k}(\mathbf{v},\mathbf{v}')\leqq k(\mathbf{v},\mathbf{v}'),\quad \mathbf{v},\mathbf{v}'\in\mathbb{R}^3,\quad \mathbf{v}\neq\mathbf{v}',\quad v\leqq v'. \tag{4.3}$$

Thus, the relations (4.2) and (4.3) give

$$\tilde{k}(\mathbf{v},\mathbf{v}')\leqq k(\mathbf{v},\mathbf{v}')+\frac{c}{r^{2-\gamma}}(1+r^2)^{1-\sigma_0/2+\sigma/2},\qquad \mathbf{v},\mathbf{v}'\in\mathbb{R}^3,\quad \mathbf{v}\neq\mathbf{v}'. \tag{4.4}$$

Due to (4.1), (2.1), and Lemma 4.2 there holds

$$\sup_{\mathbf{v}'\in\mathbb{R}^3}\int_{\mathbb{R}^3}k(\mathbf{v},\mathbf{v}')\,d\mathbf{v}<\infty,\ \sup_{\mathbf{v}\in\mathbb{R}^3}\int_{\mathbb{R}^3}k(\mathbf{v},\mathbf{v}')\,d\mathbf{v}'<\infty. \tag{4.5}$$

Finally (4.4), and (4.5) together with Lemma 4.1 show that the integral operator $\tilde{K}$ with kernel $\tilde{k}$ is bounded on $L^p(\mathbb{R}^3)$ for $1\leqq p\leqq\infty$, which was to be proved.

The first assertion of the proposition is proved in the same way. The only difference is the following. Since the case $\kappa=1$ is not excluded, Lemma 4.2 is not applicable; hence the second relation in (4.5) cannot be referred to. As a consequence, using Lemma 4.1 we obtain just boundedness of $\tilde{K}$ as a mapping from $L^1(\mathbb{R}^3)$ into itself. □

LEMMA 4.5. *Let* $\mathbf{\Gamma}$ *be bounded on* $\mathbb{R}^3\times[0,T]$, *where* $T>0$. *Then there exists a constant* $c>1$ (*depending on T*) *such that the following assertion holds*: *For every* $t\in[0,T]$ *and every* $\mathbf{u},\mathbf{y}\in\mathbb{R}^3$,

$$\frac{1}{c}(1+u^2)\leqq 1+|\mathbf{v}(\mathbf{u},\mathbf{y},t)|^2\leqq c(1+u^2).$$

Proof. Differentiate v^2 along the characteristics and use Gronwall's lemma. □

Notation. Let $L^p_{\sigma}(\mathbb{R}^3\times\mathbb{R}^3)=L^{p;p}_{\sigma}(\mathbb{R}^3\times\mathbb{R}^3)$ with norms $\|g(\cdot,\cdot)\|_{p,\sigma}=\|g(\cdot,\cdot)\|_{p,\sigma;p}$ (cf. § 3).

THEOREM 4.6. *Suppose that* $\tilde{\Gamma}_0<\infty$, $-1<\gamma\leqq 0$, $1\leqq p\leqq\infty$ *if* $\kappa\neq 1$, $p=1$ *if* $\kappa=1$. *Furthermore, let* $\operatorname{ess-sup}_{\mathbf{x}\in\mathbb{R}^3,t\in[0,T]}\|F(\cdot,\mathbf{x},t)\|_{L^{\infty}_{\sigma_0}(\mathbb{R}^3)}<\infty$ *for some* $\sigma_0>3$, $T>0$, *and let*

$\varphi \in L^p_{\sigma,+}(\mathbb{R}^3\times\mathbb{R}^3)$ *with* $0\leqq\sigma<\sigma_0-3-\gamma$. *Then there exists a unique positive mild solution* f *of the problem* (0.6) *such that*

$$\|f(\cdot,\cdot,t)\|_{p,\sigma}\leqq C_T\|\varphi\|_{p,\sigma},\qquad t\in[0,T],$$

C_T *being a constant* (*depending on* T).

Proof. Existence. Consider the sequence $\{f_n\}_{n=0}^{\infty}$ of functions defined on $\mathbb{R}^3\times\mathbb{R}^3\times\mathbb{R}_+$ recursively by (2.3). Clearly, by a Minkowsky inequality,

$$\begin{aligned}\|f^{\#}_{n+1}(\cdot,\cdot,t)\|_{p,\sigma}&\leqq\|\varphi\|_{p,\sigma}\\ &+\int_0^t\|(Kf_n)^{\#}(\cdot,\cdot,s)\|_{p,\sigma}\,ds,\qquad t\in[0,T],\quad n=0,1,\ldots.\end{aligned}\tag{4.6}$$

In view of Lemma 4.5 it is easy to verify that for every function g defined on $\mathbb{R}^3\times\mathbb{R}^3\times[0,T]$ such that $g(\cdot,\cdot,s)\in L^p_\sigma(\mathbb{R}^3\times\mathbb{R}^3)$, $s\in[0,T]$, there holds

$$c^{-\sigma/2}\|g(\cdot,\cdot,s)\|_{p,\sigma}\leqq\|g^{\#}(\cdot,\cdot,s)\|_{p,\sigma}\leqq c^{\sigma/2}\|g(\cdot,\cdot,s)\|_{p,\sigma},\qquad s\in[0,T],$$

where $c>1$ is the constant figuring in Lemma 4.5. Using this fact, the relation (4.6) implies

$$\begin{aligned}\|f_{n+1}(\cdot,\cdot,t)\|_{p,\sigma}&\leqq c^{\sigma/2}\|\varphi\|_{p,\sigma}\\ &+c^{\sigma}\int_0^t\|(Kf_n)(\cdot,\cdot,s)\|_{p,\sigma}\,ds,\qquad t\in[0,T],\quad n=0,1,\ldots.\end{aligned}$$

By Proposition 4.4 the operator $K: L^p_\sigma\to L^p_\sigma$ is bounded so that

$$\begin{aligned}\|f_{n+1}(\cdot,\cdot,t)\|_{p,\sigma}&\leqq c^{\sigma/2}\|\varphi\|_{p,\sigma}\\ &+c^{\sigma}k_{p,\sigma}\int_0^t\|f_n(\cdot,\cdot,s)\|_{p,\sigma}\,ds,\qquad t\in[0,T],\quad n=0,1,\ldots,\end{aligned}$$

where $k_{p,\sigma}$ denotes the norm of K. Using a lemma of Gronwall type, we find that

$$\|f_n(\cdot,\cdot,t)\|_{p,\sigma}\leqq C_T\|\varphi\|_{p,\sigma},\qquad t\in[0,T],\quad n=0,1,\ldots.\tag{4.7}$$

Now set $f(\mathbf{v},\mathbf{x},t)=\lim_{n\to\infty}f_n(\mathbf{v},\mathbf{x},t)$, $(\mathbf{v},\mathbf{x})\in\mathbb{R}^3\times\mathbb{R}^3$, $t\in[0,T]$. Then f is nonnegative (since all the iterates f_n are nonnegative), and by Levi's monotone convergence theorem f is measurable and, also using (4.7), we have

$$\begin{aligned}\int_{\mathbb{R}^3\times\mathbb{R}^3}(f(\mathbf{v},\mathbf{x},t)(1+v^2)^{\sigma/2})^p\,d\mathbf{v}\,d\mathbf{x}&=\lim_{n\to\infty}\int_{\mathbb{R}^3\times\mathbb{R}^3}(f_n(\mathbf{v},\mathbf{x},t)(1+v^2)^{\sigma/2})^p\,d\mathbf{v}\,d\mathbf{x}\\ &\leqq(C_T\|\varphi\|_{p,\sigma})^p,\qquad t\in[0,T].\end{aligned}$$

Moreover, letting $n\to\infty$ in (2.4) we obtain a mild solution of the problem (0.6).

Uniqueness. Now suppose that there are two mild solutions g and h. Then the difference $g-h$ is a mild solution of the problem (0.6) with $\varphi\equiv0$. Hence,

$$\|(g-h)(\cdot,\cdot,t)\|_{p,\sigma}\leqq k_{p,\sigma}\int_0^t\|(g-h)(\cdot,\cdot,s)\|_{p,\sigma}\,ds,\qquad t\in[0,T]$$

and an application of Gronwall's lemma gives $g(\cdot,\cdot,t)=h(\cdot,\cdot,t)$ in $L^p_\sigma(\mathbb{R}^3\times\mathbb{R}^3)$ for every $t\in[0,T]$. □

Next, we show that the gain collision operator K is bounded on $L^1_\sigma\cap L^p_\sigma$, including the case $\kappa=1$. This leads to the existence of solutions to (0.6) in $L^1_\sigma\cap L^p_\sigma$, see Theorem 4.9.

DEFINITION. The norm on $L^1_\sigma\cap L^p_\sigma$ is $\|\cdot\|_{1\cap p,\sigma}$.

PROPOSITION 4.7. *Let* $-1<\gamma\leqq0$, *and let* $F\in L^1_{\sigma,+}(\mathbb{R}^3)\cap L^\infty_\sigma(\mathbb{R}^3)$ *for some* $\sigma\geqq0$. *Then K is a bounded linear operator from* $L^1_\sigma(\mathbb{R}^3)$ *into itself. The norm of K is majorized by a multiple of* $\|F\|_{1\cap\infty,\sigma}$.

Proof. Using the kinetic energy conservation law and splitting the domain of integration we find after some calculations that

$$\int_{\mathbb{R}^3} |Kf(\mathbf{v})(1+v^2)^{\sigma/2}|\, d\mathbf{v} \leqq \text{const.}\left(\frac{4\pi}{3+\gamma}\|F\|_{\infty,\sigma}\|f\|_{1,\sigma} + \|F\|_{1,\sigma}\|f\|_{1,\sigma}\right). \qquad \square$$

PROPOSITION 4.8. *Suppose that* $-1<\gamma\leqq 0$, $F\in L^1_{\sigma_0,+}(\mathbb{R}^3)\cap L^\infty_{2+\sigma_0}(\mathbb{R}^3)$ *for some* $\sigma_0>0$, $0\leqq\sigma\leqq\sigma_0$, *and that* $1\leqq p\leqq\infty$. *Then K is a bounded operator from* $L^1_\sigma(\mathbb{R}^3)\cap L^p_\sigma(\mathbb{R}^3)$ *into itself. The norm of K is majorized by a multiple of* $\|F\|_{1\cap\infty,2+\sigma_0}$.

Proof. By Lemma 3.6, and Proposition 4.7, using linear interpolation (cf. the proof of Theorem 3.8) we obtain the asserted result. $\square$

The following theorem can now be proved in the same manner as Theorem 4.6.

THEOREM 4.9. *Suppose that* $-1<\gamma\leqq 0$, $\tilde{\Gamma}_0<\infty$, $1\leqq p\leqq\infty$, *and that* ess-$\sup_{\mathbf{x}\in\mathbb{R}^3, t\in[0,T]}\|F(\cdot,\mathbf{x},t)\|_{1\cap\infty,2+\sigma_0}<\infty$ *for some* $\sigma_0>0$, $T>0$. *If* $\varphi\in L^1_{\sigma,+}(\mathbb{R}^3\times\mathbb{R}^3)\cap L^p_\sigma(\mathbb{R}^3\times\mathbb{R}^3)$, *where* $0\leqq\sigma\leqq\sigma_0$, *then there exists a unique nonnegative mild solution f of the problem* (0.6) *such that*

$$\|f(\cdot,\cdot,t)\|_{1\cap p,\sigma}\leqq C_T\|\varphi\|_{1\cap p,\sigma}, \qquad t\in[0,T].$$

Here C_T *is a constant* (*depending on* T).

Remark. A similar result can also be obtained by combining the results of Theorem 2.4 and Theorem 4.6.

REFERENCES

[Ar] L. ARKERYD, *On the Boltzmann equation*, Arch. Rational Mech. Anal., 45 (1972), pp. 1-34.

[BL] J. BERGH AND J. LÖFSTRÖM, *Interpolation Spaces*, Springer-Verlag, Berlin, 1976.

[Ca] T. CARLEMAN, *Problèmes mathématiques dans la théorie cinétique des gaz*, Almqvist & Wiksells, Uppsala, 1957.

[Ce] C. CERCIGNANI, *The Boltzmann Equation and Its Applications*, Springer-Verlag, New York, 1988.

[CGP] F. CHVÁLA, T. GUSTAFSSON, AND R. PETTERSSON, *On moments of solutions to the linear Boltzmann equation with external electromagnetic force*, Dept. of Mathematics, Chalmers Univ. of Technology Gōtehorg, 1989, preprint.

[Ch] F. CHVÁLA, *On the Boltzmann collision term and related operators*, Acta Tech. ČSAV, 30 (1985), pp. 327-364.

[Ed] R. E. EDWARDS, *Functional Analysis. Theory and Applications.* Holt, Rinehart, and Winston, New York, 1965.

[GMP] W. GREENBERG, C. VAN DER MEE, AND V. PROTOPOPESCU, *Boundary Value Problems in Abstract Kinetic Theory*, Birkhäuser Verlag, Basel, Switzerland, 1987.

[Gr] H. GRAD, *Asymptotic Theory of the Boltzmann Equation*, II. *Rarefied Gas Dynamics*, Vol. 1, J. A. Laurmann, ed., Academic Press, New York, 1963, pp. 26-59.

[Gu1] T. GUSTAFSSON, *L^p-estimates for the nonlinear spatially homogeneous Boltzmann equation*, Arch. Rational Mech. Anal., 92 (1986), pp. 23-57.

[Gu2] ———, *Global L^p-Properties for the Spatially Homogeneous Boltzmann Equation*, Arch. Rational Mech. Anal., 103 (1988), pp. 1-38.

[KS] S. KANIEL AND M. SHINBROT, *The Boltzmann equation*, I. *Uniqueness and local existence*, Comm. Math. Phys., 58 (1978), pp. 65-84.

[Ku] J. KURZWEIL, *Ordinary Differential Equations*, Elsevier, Amsterdam, 1986.

[Mo] F. A. MOLINET, *Existence, uniqueness and properties of the solutions of the Boltzmann kinetic equation for a weakly ionized gas*, J. Math. Phys., 18 (1977), pp. 984-996.

[P1] R. PETTERSSON, *Existence theorems for the linear, space-inhomogeneous transport equation*, IMA J. Appl. Math., 30 (1983), pp. 81-105.

[P2] ———, *On solutions and higher moments for the linear Boltzmann equation with infinite-range forces*, IMA J. Appl. Math., 38 (1987), pp. 151-166.

[PL] R. J. DI PERNA AND P. L. LIONS, *On the Cauchy problem for Boltzmann equations, global existence and weak stability*, Ann. of Math., 130 (1989), pp. 321-366.

[TM] C. TRUESDELL AND R. G. MUNCASTER, *Fundamentals of Maxwell's Kinetic Theory of a Simple Monoatomic Gas*, Academic Press, New York, 1980.

SIAM J. MATH. ANAL.
Vol. 24, No. 3, pp. 603–617, May 1993

ON THE AMPLITUDE EQUATIONS ARISING AT THE ONSET OF THE OSCILLATORY INSTABILITY IN PATTERN FORMATION*

JOSÉ M. VEGA†

Abstract. A well-known system of two amplitude equations is considered that describes the weakly nonlinear evolution of many nonequilibrium systems at the onset of the so-called oscillatory instability. Those equations depend on a small parameter, ε, that is a ratio between two distinguished spatial scales. In the limit $\varepsilon \to 0$, a simpler asymptotic model is obtained that consists of two complex cubic Ginzburg–Landau equations, coupled only by spatially averaged terms.

Key words. pattern formation, oscillatory instability, amplitude equations, Ginzburg–Landau equations

AMS subject classifications. 35K55, 35B25, 35B32, 35B35

1. Introduction and formulation. This paper deals with the following parabolic problem:

$$A_t = (1+i\alpha)A_{xx} - \varepsilon-1A_x + \mu A - (1+i\beta)A|A|^2 + (\gamma+i\delta)A|B|^2, \tag{1.1}$$

$$B_t = (1+i\alpha)B_{xx} + \varepsilon-1B_x + \mu B - (1+i\beta)B|B|^2 + (\gamma+i\delta)B|A|^2, \tag{1.2}$$

$$A(x+1,t) = A(x,t)\exp(-ia), \quad B(x+1,t) = B(x,t)\exp(ia) \text{ if } x \in \mathbb{R},\ t > 0, \tag{1.3}$$

$$A(x,0) = A_0(x), \quad B(x,0) = B_0(x) \quad \text{if } x \in \mathbb{R}. \tag{1.4}$$

Here A and B are appropriately smooth complex-valued functions of the spatial variable x and the time t; i is the imaginary unit; the real parameters α, β, γ, δ, μ, and a are such that

$$\mu > 0, \quad \gamma < 1, \quad \text{and} \quad 0 \le a < 2\pi, \tag{1.5}$$

and the real parameter $\varepsilon > 0$ is assumed to be small. The functions A_0 and B_0 appearing in the initial conditions (1.4) are appropriately smooth. Then, as we shall see in §3, the problem (1.1)–(1.4) is well posed and its solution is defined in $0 < t < \infty$. In addition, we shall consider the problem (1.1)–(1.4) under one of the following additional conditions:

$$A(x,t) = -B(-x,t) \quad \text{if } x \in \mathbb{R},\ t \ge 0, \tag{1.6}$$

$$A(x,t) = B(-x,t) \quad \text{if } x \in \mathbb{R},\ t \ge 0. \tag{1.7}$$

*Received by the editors March 4, 1991; accepted for publication (in revised form) September 2, 1992. This research was partially supported by Dirección General de Investigación Científica y Técnica grant PB-90-0271.

†Escuela Técnica Superior de Ingenieros Aeronáuticos, Universidad Politécnica de Madrid, 28040 Madrid, Spain.

As we shall see in §3 each of these conditions is compatible with (1.1)–(1.4) in the following sense: if a solution of (1.1)–(1.4) satisfies that condition for $t = 0$, then the condition holds for all $t > 0$.

Equations (1.1), (1.2) appear as a normal form in a weakly nonlinear analysis of nonequilibrium systems at the onset of the so-called *oscillatory instability*; see [1]–[4]. In particular, they have been used in the analysis of wave dynamics in doubly diffusive and binary fluid convection (see [1], [5], [6]), and some of their solutions have been seen to describe qualitatively the results of some experiments (see [7]–[11]).

Let us now explain briefly how these equations are obtained and where conditions (1.3), (1.6), and (1.7) come from. We consider a system of PDEs of the form

$$\frac{\partial u}{\partial T} = G\left(\frac{\partial^2}{\partial X^2}, u; R\right) \quad \text{in } -\infty < X < \infty,\ T > 0, \tag{1.8}$$

where $u = (u_1, \ldots, u_N)$ is a function of the space variable X, and the time variable T, R is a control parameter, and G is a nonlinear differential operator (invariant under space translations, $X \to X + c$, and reflection, $X \to -X$) such that $G(\partial^2/\partial X^2, 0; R) \equiv 0$ (the uniform state $u \equiv 0$ is a stationary solution of (1.8) for all R).

Let $\varphi(\omega, k^2, R) = 0$ (ω = growth rate, k = wave number) be the complex dispersion relation of the linearized problem about $u = 0$. We assume that the growth rate of the most unstable modes are of the form

$$\omega = \pm i\Omega + c_{1\pm}R \mp c_2 k_0(k - k_0) - c_{3\pm}(k - k_0)^2 + o(|R|) + o(|k - k_0|^2),$$

$$\omega = \pm i\Omega + c_{1\pm}R \pm c_2 k_0(k + k_0) - c_{3\pm}(k + k_0)^2 + o(|R|) + o(|k + k_0|^2),$$

as $R \to 0$ and $k \to \pm k_0$, while $\operatorname{Re}\omega < 0$ otherwise. Here the coefficients Ω and k_0 are real, while $c_{1\pm}$, c_2 and $c_{3\pm}$ are complex, and such that $c_{1-} = \bar{c}_{1+}$, $c_{3-} = \bar{c}_{3+}$ (overbars stand for the complex conjugate) and

$$\Omega > 0, \quad k_0 > 0, \quad \operatorname{Re} c_{1+} > 0, \quad \operatorname{Re} c_2 = 0, \quad \operatorname{Re} c_{3+} > 0. \tag{1.9}$$

Notice that the corresponding neutrally stable modes at $R = 0$ are of the form

$$u = U_0[A\exp(i\Omega T + ik_0 X) + B\exp(i\Omega T - ik_0 X)] + \text{c.c.}$$

for a certain eigenvector $U_0 \in \mathbb{C}^N$, where the complex amplitudes A and B are arbitrary and c.c. stands for the complex conjugate. That mode is the superposition of two wavetrains that are travelling in opposite directions. The weakly nonlinear stability of those pairs of waves, as $R \to 0$, is analyzed by appropriately scaling the complex amplitudes and allowing them to depend on the slow time and space variables $t = \varepsilon^2 T$ and $x = \varepsilon X$, where the small parameter $\varepsilon > 0$ is defined by $\varepsilon^2 = R$. If the ansatz

$$u = \varepsilon U_0[A(x,t)\exp(i\Omega T + ik_0 X) + B(x,t)\exp(i\Omega T - ik_0 X)] + \text{c.c.} + O(\varepsilon^2) \tag{1.10}$$

is inserted into (1.8) and the appropriate solvability conditions are applied at the orders $O(\varepsilon^2)$ and $O(\varepsilon^3)$, then the following evolution equations are obtained (at the order $O(\varepsilon^3)$):

$$A_t = c_{3+}A_{xx} - \left(\frac{c_2 k_0}{i\varepsilon}\right)A_x + c_{1+}A + c_4 A|A|^2 + c_5 A|B|^2, \tag{1.11}$$

$$B_t = c_{3+}B_{xx} + \left(\frac{c_2k_0}{i\varepsilon}\right)B_x + c_{1+}B + c_4B|B|^2 + c_5B|A|^2, \tag{1.12}$$

where the coefficients c_{1+}, c_2, and c_{3+} are as defined above, and c_4 and c_5 depend on the leading nonlinear terms of (1.8). Equations (1.11), (1.12) may be obtained quite directly by symmetry considerations (see [1]), but the actual values of the coefficients c_4 and c_5 must be obtained by the process described above (which usually leads to quite tedious calculations, even with the help of symbolic algebra). Notice that the coefficient $(c_2k_0/i\varepsilon)$ is real and large as $\varepsilon \to 0$ if $c_2 \neq 0$ as we are assuming (the assumption $c_2 = 0$ would restrict the analysis to a codimension-two point of the parameter space of (1.8)). In addition to (1.9), the following supercriticality assumption will be made:

$$\operatorname{Re} c_4 < 0, \qquad \operatorname{Re}(c_4 + c_5) < 0. \tag{1.13}$$

In fact, if (1.13) does not hold, then (1.11), (1.12) possess solutions that either are unbounded as $t \to \infty$ or blow up in a finite time, as is readily seen by considering spatially uniform solutions.

Now, let us impose the following spatial periodicity condition on the solutions of (1.8):

$$u(X + L, T) \equiv u(X, T), \tag{1.14}$$

where the period L depends on ε in such a way that $\varepsilon L \to l > 0$ and $2\pi\,\mathrm{fract}\,(k_0L/2\pi) \to d$ as $\varepsilon \to 0$ (fract(z) stands for the fractionary part of the real z). Then if condition (1.14) is imposed on (1.10), the following conditions are obtained (to the leading order):

$$A(x + l, t) \equiv A(x, t)\exp(-id), \quad B(x + l, t) \equiv B(x, t)\exp(id). \tag{1.15}$$

Observe that the conditions (1.14) are appropriate to analyze the problem (1.8) in a finite interval, $0 < X < L/2$, if the boundary conditions at $X = 0$ and $L/2$ are either of the Dirichlet ($u = 0$) or Neumann ($u_X = 0$) type. This is readily seen when u is extended to $-\infty < X < \infty$ by means of the appropriate reflexion principle:

$$u(X + mL/2, T) \equiv -u(-X + mL/2, T) \quad \text{(Dirichlet), or} \tag{1.16}$$

$$u(X + mL/2, T) \equiv u(-X + mL/2, T) \quad \text{(Neumann)} \tag{1.17}$$

for each integer m. When (1.16) or (1.17) are imposed on (1.10), then (to leading order) we obtain (1.15) and

$$A(x, t) \equiv -B(-x, t) \text{ (Dirichlet), or} \tag{1.18}$$

$$A(x, t) \equiv B(-x, t) \text{ (Neumann).} \tag{1.19}$$

Now, by applying in (1.11), (1.12), (1.15) an appropriate transformation of the type $A(x, t) \to A(x, t)\exp(iet)$, $B(x, t) \to B(x, t)\exp(iet)$ (to make real the coefficients of A and B in (1.11) and (1.12), respectively), and rescaling the variables A, B, x, and t, and the small parameter ε, we obtain (1.1)–(1.3); conditions (1.5) readily follow from

(1.13). By the same process, the additional condition (1.18) (respectively, (1.19)) leads to (1.6) (respectively, (1.7)).

Therefore, the problem (1.1)–(1.4) seems to be appropriate to analyze spatially periodic patterns of (1.8) near the instability limit. Pattern formation in finite intervals, under homogeneous Dirichlet or Neumann boundary conditions, is analyzed by imposing (1.6) or (1.7).

The main object of this paper is to obtain rigorously the following approximate model (of (1.1)–(1.4)) in the limit $\varepsilon \to 0$: $A(x,t) \simeq V(x - t/\varepsilon, t)$, $B(x,t) \simeq W(x + t/\varepsilon, t)$, where the functions $(y,t) \to V(y,t)$ and $(z,t) \to W(z,t)$ satisfy, in first approximation,

$$V_t = (1 + i\alpha)V_{yy} + \mu V - (1 + i\beta)V|V|^2 + (\gamma + i\delta)V\langle|W|^2\rangle, \tag{1.20}$$

$$W_t = (1 + i\alpha)W_{zz} + \mu W - (1 + i\beta)W|W|^2 + (\gamma + i\delta)W\langle|V|^2\rangle, \tag{1.21}$$

$$V(y + 1, t) = V(y,t)\exp(-ia), \quad W(z + 1, t) = W(z,t)\exp(ia), \tag{1.22}$$

$$V(y,0) = A_0(y), \quad W(z,0) = B_0(z), \tag{1.23}$$

if $y, z \in \mathbb{R}$, $t > 0$. Here, the spatial average $\langle\cdot\rangle$ is defined by

$$\langle f\rangle = \int_0^1 f(\xi,t)d\xi \tag{1.24}$$

for each function $f \in C(\mathbb{R} \times [0,\infty[)$ that is periodic, of period 1, in its first argument (as $|V|^2$ and $|W|^2$ are; see (1.22)). In addition, we shall see that if A and B satisfy one of the additional conditions, (1.6) or (1.7), then $\langle|W|^2\rangle = \langle|V|^2\rangle$ for all $t \geq 0$, and the model (1.20)–(1.23) may be further simplified to

$$V_t = (1 + i\alpha)V_{yy} + \mu V - (1 + i\beta)V|V|^2 + (\gamma + i\delta)V\langle|V|^2\rangle, \tag{1.25}$$

$$V(y + 1, t) = V(y,t)\exp(-ia), \ V(y,0) = A_0(y) \quad \text{if } y \in \mathbb{R}, \ t > 0. \tag{1.26}$$

These two models were first obtained, independently, by Knobloch and De Luca [12] and by Alvarez-Pereira and Vega [13]. In both cases a formal derivation was made, by means of perturbation techniques. In fact, in [13] the problem under consideration was not of the type (1.8), but (1.20)–(1.23) and (1.25)–(1.26) appeared in a weakly nonlinear stability analysis, in two space dimensions, of travelling plane wave-fronts in a reaction-diffusion system arising in combustion theory; that problem is essentially more involved than (1.8), and the models (1.20)–(1.23) and (1.25)–(1.26) were obtained in a particular limit.

The paper is organized as follows. For convenience, a formal derivation (by means of a two-time scales method) of (1.20)–(1.23) is given in §2. Section 3 includes some preliminaries concerning the model (1.1)–(1.4), a rigorous derivation of the approximate models and some basic properties of these models. Finally, in §4 we describe further properties of the approximate models, and make some conjectures.

2. A formal derivation of the model (1.20)–(1.23). Here we give a formal derivation of model (1.20)–(1.23) by means of perturbation techniques (the second asymptotic model is easily obtained from the first one, as it will be seen in §3). That derivation will give the key idea for the rigorous analysis of next section. In addition, it will provide some insight into the nature of the limit $\varepsilon \to 0$, by explaining why it is natural to expect the appearance of the spatial averaged terms in the approximate models.

As $\varepsilon \to 0$, there are two obvious time scales in (1.1), (1.2):

$$t \sim 1 \quad \text{and} \quad \tau \equiv t/\varepsilon \sim 1. \tag{2.1}$$

The latter comes when balancing A_t with $-\varepsilon^{-1}A_x$ in (1.1) (or B_t with $\varepsilon^{-1}B_x$ in (1.2)). Then we shall seek the expansions

$$A(x,t) = \tilde{A}_0(x,\tau,t) + \varepsilon\tilde{A}_1(x,\tau,t) + \cdots, \quad B(x,t) = \tilde{B}_0(x,\tau,t) + \varepsilon\tilde{B}_1(x,\tau,t) + \cdots.$$

When these expansions are inserted into (1.1)–(1.3), and the coefficients of ε^0 and ε^1 are set to zero, the following problems are obtained:

$$\tilde{A}_{0\tau} + \tilde{A}_{0x} = 0, \quad \tilde{B}_{0\tau} - \tilde{B}_{0x} = 0, \tag{2.2}$$

$$\tilde{A}_{1\tau} + \tilde{A}_{1x} = -\tilde{A}_{0t} + (1+i\alpha)\tilde{A}_{0xx} + \tilde{A}_0\left[\mu - (1+i\beta)|\tilde{A}_0|^2 + (\gamma+i\delta)|\tilde{B}_0|^2\right], \tag{2.3}$$

$$\tilde{B}_{1\tau} - \tilde{B}_{1x} = -\tilde{B}_{0t} + (1+i\alpha)\tilde{B}_{0xx} + \tilde{B}_0\left[\mu - (1+i\beta)|\tilde{B}_0|^2 + (\gamma+i\delta)|\tilde{A}_0|^2\right], \tag{2.4}$$

$$\tilde{A}_j(x+1,\tau,t) \equiv \tilde{A}_j(x,\tau,t)\exp(-ia), \quad \tilde{B}_j(x+1,\tau,t) \equiv B_j(x,\tau,t)\exp(ia), \tag{2.5}$$

$$\tilde{A}_0(x,0,0) \equiv A_0(x), \quad \tilde{B}_0(x,0,0) \equiv B_0(x), \quad \tilde{A}_1(x,0,0) \equiv \tilde{B}_1(x) \equiv 0 \tag{2.6}$$

for $j = 0$ and 1. The wave equations (2.2) readily yield $\tilde{A}_0(x,\tau,t) \equiv V_0(x-\tau,t)$, $\tilde{B}_0(x,\tau,t) \equiv W_0(x+\tau,t)$, for some functions $(y,t) \to V_0(y,t)$ and $(z,t) \to W_0(z,t)$ that satisfy (1.22), (1.23) (see (2.5), (2.6)) and are otherwise arbitrary (at the moment). Then the solution of (2.3), (2.4) may be found in close-form

$$\begin{aligned}\tilde{A}_1(x,\tau,t) = V_1(x-\tau,t) &+ \tau\left[-V_{0t} + (1+i\alpha)V_{0yy} + V_0\left(\mu - (1+i\beta)|V_0|^2\right)\right]\\ &+ \left(\frac{\gamma+i\delta}{2}\right)V_0\int_0^{\tau+x}|W_0(z,t)|^2dz,\end{aligned} \tag{2.7}$$

$$\begin{aligned}\tilde{B}_1(x,\tau,t) = W_1(x+\tau,t) &+ \tau\left[-W_{0t} + (1+i\alpha)W_{0zz} + W_0\left(\mu - (1+i\beta)|W_0|^2\right)\right]\\ &+ \left(\frac{\gamma+i\delta}{2}\right)W_0\int_0^{\tau-x}|V_0(y,t)|^2dy\end{aligned} \tag{2.8}$$

for some functions V_1 and W_1 (to be determined at later stages).

Now, by eliminating secular terms in the fast time scale, i.e., by requiring the right-hand side of (2.7) to be bounded as $\tau \to \infty$ ($t =$ constant in this time scale) we readily obtain

$$-V_{0t} + (1+i\alpha)V_{0yy} + V_0\left[\mu - (1+i\beta)|V_0|^2\right] \\ + (\gamma + i\delta)V_0 \lim_{\tau\to\infty}(2\tau)^{-1}\int_0^{\tau+x} |W_0(z,t)|^2 dz = 0.$$

Notice that for each fixed value of $y = x - \tau$, the first three terms in the left-hand side of this equation are constant (recall that $t =$ constant at this time stage), while the last one is of the form $(\gamma + i\delta)V_0 \lim_{\tau\to\infty}(2\tau)^{-1}\int_0^{2\tau+y} |W_0(z,t)|^2 dz$, and since $|W_0(z+1,t)|^2 \equiv |W_0(z,t)|^2$ (see (1.22)), we have $(2\tau)^{-1}\int_0^{2\tau+y} |W_0(z,t)|^2 dz = \langle |W_0|^2\rangle + O(\tau^{-1})$ as $\tau \to \infty$, where the spatial average $\langle\cdot\rangle$ was defined in (1.24). Then V_0 and W_0 satisfy (1.20). Equation (1.21) is obtained in a completely similar way, by eliminating secular terms in (2.8), and the derivation of (1.20)–(1.23) is complete.

Observe that, to the leading order, in the fast time scale ($\tau \sim 1$), the amplitudes A and B satisfy the wave equations (2.2), whose solution represents two waves that are travelling in opposite directions. If one moves in a reference frame attached to the wave associated with A (i.e., if $x - \tau =$ constant), then the wave associated with B is seen to travel at a speed 2 in the fast time scale, or at a speed $2/\varepsilon$ ($\to \infty$ as $\varepsilon \to 0$) in the slow time scale. Therefore, if $t \sim 1$, then the spatial structure of B is not appreciated from the reference frame moving with A; only the spatial mean value of $|B|^2$ over a period is seen. This explains the appearance of the spatial averages in the asymptotic equations, and suggests the main argument in the proof of Theorem 3.4 below.

3. Main results concerning the models (1.1)–(1.4), (1.20)–(1.23), and (1.25)–(1.26).

Here we first consider some preliminary properties of the model (1.1)–(1.4). Then, the approximate models (1.20)–(1.23) and (1.25)–(1.27) are rigorously derived. Finally, some basic properties of the approximate models are given.

3.1. The model (1.1)–(1.4).

In order to prove that the model (1.1)–(1.4) is well posed we could modify appropriately an abstract result by Ghidaglia [14] that was used be Temam [15] to prove the well-posedness of the standard complex cubic Ginzburg–Landau equation in finite domains with standard boundary conditions. Nevertheless, for the sake of brevity, we shall follow a more direct approach, based on a classical result by Henry [16].

THEOREM 3.1. *If $\gamma < 1$ and if the functions A_0 and B_0 belong to* (*the complexified space of*) $C^{2+\alpha}([0,1])$, *for some* $\alpha \in]0,1[$, *and satisfy* (1.3), *then* (1.1)–(1.4) *possess a unique classical solution,* $(x,t) \to (A(x,t), B(x,t))$, *such that* $A, B \in C^{2+\alpha,1+\alpha/2}([0,1] \times [0,T])$ *for all* $T > 0$. *In addition,* $A, B \in C^{2k,k}([0,1]\times[T_1,T_2])$ *whenever* $0 < T_1 < T_2$ *and* $k \geq 0$.

Proof. We first state (1.1)–(1.4) in an appropriate abstract setting (after decomplexification and restriction to the bounded interval $0 \leq x \leq 1$, with the appropriate boundary conditions at $x = 0, 1$, obtained from (1.3)) as

$$\frac{du}{dt} + Lu = f(u) \quad \text{if } t > 0, \qquad u(0) = u_0 \in D(L), \tag{3.1}$$

where the linear operator $L : D(L) \to X \equiv L_2(]0,1[)^4$ is defined in such a way that $f : X \to X$ is a substitution perator, and

$$D(L) = \{u \in H_2(]0,1[)^4 : u \text{ satisfies the boundary conditions at } x = 0,1\}.$$

Then the operator L is sectorial in X, and f is locally Lipschitzian and maps bounded sets into bounded sets. Then, by applying [17, p. 55, Thm. 3.3.4], it follows that (3.1) possesses a unique mild solution, $u \in C([0,T[, X)$, in a maximal interval of existence, $0 \le t < T$, and one of the following alternatives holds: either (i) $T = \infty$, or (ii) $T < \infty$ and there is a sequence, $\{t_m\} \subset \mathbb{R}$, such that $t_m \to T^-$ and $\|u(t_m)\|_X \to \infty$. But since $\gamma < 1$, according to the result in Lemma 3.3 below (that may be seen as an a priori estimate on (3.1)) the alternative (ii) cannot hold. Then $T = \infty$, and we obtain global existence and uniqueness of mild solutions. Further regularity of u is obtained as usual, by means of parabolic estimates [17, VII-10] (those estimates are readily extended to apply for boundary conditions such as (1.3)) and imbedding theorems [17, II-3]. □

Now we prove that the additional conditions (1.6) and (1.7) are compatible with the model (1.1)–(1.4).

LEMMA 3.2. *If a classical solution of* (1.1)–(1.4) *satisfies* (1.6) (*respectively,* (1.7)) *at* $t = 0$, *then* (1.6) (*respectively,* (1.7)) *holds for all* $t > 0$.

Proof. Let (A, B) be a classical solution of (1.1)–(1.4) satisfying (1.6) (if it satisfies (1.7) the argument is similar) at $t = 0$, and let the pair of functions A_1 and B_1 be defined by $A_1(x,t) \equiv -B(-x,t)$ and $B_1(x,t) \equiv -A(-x,t)$. The pair (A_1, B_1) satisfies (1.1)–(1.3), and since (1.6) holds at $t = 0$, $A_1(x,0) = A(x,0)$ and $B_1(x,0) = B(x,0)$ for all $x \in \mathbb{R}$. Then, by uniqueness of (1.1)–(1.4), $A_1(x,t) = A(x,t)$ and $B_1(x,t) = B(x,t)$ for all $x \in \mathbb{R}$ and all $t > 0$, i.e., condition (1.6) holds for all $t > 0$, as stated. □

3.2. A rigorous derivation of the approximate models. As suggested by the formal derivation in §2, the approximate models will be obtained, in Theorem 3.4, by means of an averaging method. To apply that method we need some estimates on (1.1)–(1.4) that hold uniformly as $\varepsilon \to 0$ (see Lemma 3.5 below). Two basic inequalities that will be used systematically in the proof of Lemma 3.3 are first given.

If $u \in C^1([0,1] \times \mathbb{R})$ *is such that* $\int_0^1 u(x)dx = 0$, *then*

$$\|u\|_{L_q(]0,1[)} \le k_{qr}\|u_x\|^{\alpha}_{L_r(]0,1[)} \cdot \|u\|^{1-\alpha}_{L_r(]0,1[)}, \tag{3.2}$$

whenever $q > r \ge 1$, *where* $\alpha = 1/r - 1/q$, *and the constant* k_{qr} *depends only on* q *and* r (*it is independent of* u). *If* $v \in C^1([0,1], \mathbb{C})$ *and* $q > 1$, *then*

$$\|v\|^2_{L_{2q}(]0,1[)} \le \|v\|^2_{L_2(]0,1[)} + 2k_{q1}\|v\|^{1+1/q}_{L_2(]0,1[)} \cdot \|v_x\|^{1-1/q}_{L_2(]0,1[)}, \tag{3.3}$$

where the constant k_{q1} *is defined as in* (3.2).

The inequality (3.2) is given in, e.g., [17, pp. 62–63], while (3.3) is obtained by applying (3.2) (and Hölder's inequality) to the function $x \to u(x) \equiv |v(x)|^2 - \|v\|^2_{L_2(]0,1[)}$.

We now give some uniform estimates on the solutions of (1.1)–(1.4). For the sake of brevity, we shall not try to obtain the best possible values of the bounding constants K_0, T, and C_0. Notice that these constants are independent of ε, that C_0 is also independent of the particular solution of (1.1)–(1.4), and that K_0 and T depend on it only through $\phi_j(0)$.

LEMMA 3.3. *If* $\gamma < 1$, *for each classical solution of* (1.1)–(1.4) *and for* $j = 0, 1,$ *and* 2, *let the functions* $t \to \phi_j(t)$ *be defined by*

$$\phi_j(t) \equiv \int_0^1 \left[\left| \frac{\partial^j A}{\partial x^j} \right|^2 + \left| \frac{\partial^j B}{\partial x^j} \right|^2 \right] dx. \tag{3.4}$$

Then, for $j = 0, 1,$ and 2, ϕ_j satisfies

$$\phi_j(t) \leq K_0 \quad \text{for all } t \geq 0, \quad \text{and } \phi_j(t) \leq C_0 \quad \text{for all } t \geq T, \tag{3.5}$$

where the constants K_0 and T (respectively, C_0) depend only on β, γ, δ, μ, $\phi_1(0)$, $\phi_2(0)$, and $\phi_3(0)$ (respectively, on β, γ, δ, and μ).

Proof. The estimates (3.5) will be obtained successively for ϕ_0, ϕ_1, and ϕ_2. For the sake of clarity we shall simplify the notation as follows. Every constant appearing below that depends only on β, γ, δ, μ, $\phi_1(0)$, $\phi_2(0)$, and $\phi_3(0)$ (respectively, on β, γ, δ, and μ) will be denoted always as K (respectively, as C).

Step 1. *The estimates* (3.5) *for* ϕ_0. Let us multiply (1.1) by $\bar{A}$, (1.2) by $\bar{B}$, add the resulting equations, take the real part, integrate on $0 < x < 1$, integrate by parts, use (1.3), and apply Hölder's inequality twice to obtain

$$\begin{aligned} \frac{d\phi_0}{dt} &= -2\phi_1 + 2\mu\phi_0 - 2\int_0^1 [|A|^4 + |B|^4 - 2\gamma|A|^2|B|^2]\,dx \\ &\leq -2\phi_1 + 2\mu\phi_0 - 2C\int_0^1 [|A|^4 + |B|^4]\,dx \leq -2\phi_1 + 2\mu\phi_0 - C\phi_0^2 \quad \text{if } t > 0, \end{aligned} \tag{3.6}$$

where $C = \min\{1, 1-\gamma\} > 0$. Since $\phi_1(t) \geq 0$ for all $t > 0$, the estimates (3.5) readily follow (with, e.g., $K_0 = \max\{\phi_0(0), 2\mu/C\}$, $C_0 = 4\mu/C$ and $T = (2\mu)^{-1}\log[\max\{1, 2 - 4\mu/C\phi_0(0)\}]$ if $\mu > 0$, while $K_0 = \phi_0(0)$, $C_0 = 1$, and $T = \max\{0, (\phi_0(0) - 1)/C\phi_0(0)\}$ if $\mu \leq 0$).

Step 2. *The estimates* (3.5) *for* ϕ_1. Now, we multiply (1.1) by $-\bar{A}_{xx}$, (1.2) by $-\bar{B}_{xx}$, add the resulting equations, take the real part, integrate on $0 < x < 1$, integrate by parts and use (1.3). Then, the following equation follows:

$$\begin{aligned} \frac{d\phi_1}{dt} &= -\,2\phi_2 + 2\mu\phi_1 \\ &\quad - \int_0^1 \left\{2|A|^2|A_x|^2 + 2|B|^2|B_x|^2 + [(|A|^2)_x]^2 + [(|B|^2)_x]^2\right\} dx + F_1 + F_2 \end{aligned} \tag{3.7}$$

for all $t > 0$, where

$$\begin{aligned} F_1 &= -\gamma\int_0^1 \left[(A\bar{A}_{xx} + \bar{A}A_{xx})|B|^2 + (B\bar{B}_{xx} + \bar{B}B_{xx})|A|^2\right] dx \\ &\leq 4|\gamma|\left[\int_0^1 (|A|^6 + |B|^6)\,dx \int_0^1 (|A_{xx}|^2 + |B_{xx}|^2)\,dx\right]^{1/2}, \\ F_2 &= i\int_0^1 \left[(A\bar{A}_{xx} - \bar{A}A_{xx})(\beta|A|^2 - \delta|B|^2) + (B\bar{B}_{xx} - \bar{B}B_{xx})(\beta|B|^2 - \delta|A|^2\right] dx \\ &\leq 4(|\beta| + |\delta|)\left[\int_0^1 (|A|^6 + |B|^6)\,dx \int_0^1 (|A_{xx}|^2 + |B_{xx}|^2)\,dx\right]^{1/2}, \end{aligned}$$

or, by using (3.3) (with $q = 3$), $F_1 + F_2 \leq C\phi_0(t)[\phi_0(t) + \phi_1(t)]^{1/2}\phi_2(t)^{1/2}$ if $t > 0$, where the constant C is as defined above (in fact, C depends only on β, γ, and δ). Then, (3.7) yields $d\phi_1/dt \leq -2\phi_2 + 2\mu\phi_1 + C\phi_0(\phi_0 + \phi_1)^{1/2}\phi_2^{1/2}$ if $t > 0$, and since the function $z \to f(z) \equiv -z + cz^{1/2}$ ($c \geq 0$ being a constant) satisfies $f(z) \leq c^2/4$ for all $z \geq 0$, we have

$$\frac{d\phi_1}{dt} \leq -\phi_2 + 2\mu\phi_1 + \frac{C^2\phi_0^2(\phi_0 + \phi_1)}{4} \quad \text{if } t > 0. \tag{3.8}$$

Then, if the constant $K > 0$ is such that $2\mu + C^2\phi_0(t)^2/4 \leq K$ for all $t > 0$, we multiply (3.6) by K and add the resulting equation to (3.8) to obtain $d(\phi_1 + K\phi_0)/dt \leq -K(\phi_1 + K\phi_0) + K(K + 2\mu)\phi_0 + C\phi_0^3/4$ if $t > 0$, and the first estimate (3.5) readily follows for ϕ_1. The second estimate (3.5) is obtained in a similar way, when taking into account that it holds for ϕ_0, and using an appropriate linear combination of the inequalities (3.6) and (3.8).

Step 3. *The estimates* (3.5) *for* ϕ_2. Again, we shall use the following equation, which is obtained by differentiating (1.1) and (1.2) with respect to x, multiplying the resulting equations by $-\bar{A}_{xxx}$ and $-\bar{B}_{xxx}$, respectively, adding, taking the real part, integrating on $0 < x < 1$, integrating by parts, and using (1.3):

$$\frac{d\phi_2}{dt} = -2\phi_3 + 2\mu\phi_2 + F_3 + F_4 \quad \text{if } t > 0, \tag{3.9}$$

where ϕ_3 is defined as in (3.4) and (recall that c.c. stands for the complex conjugate)

$$\begin{aligned} F_3 &= (1 + i\beta) \int_0^1 \left[\bar{A}_{xxx}\left(A|A|^2\right)_x + \bar{B}_{xxx}\left(B|B|^2\right)_x\right] dx + \text{c.c.} \\ &\leq 6|1 + \beta| \int_0^1 \left[|A|^2|A_x||A_{xxx}| + |B|^2|B_x||B_{xxx}|\right] dx, \end{aligned}$$

$$\begin{aligned} F_4 &= -(\gamma + i\delta) \int_0^1 \left[\bar{A}_{xxx}\left(A|B|^2\right)_x + \bar{B}_{xxx}\left(B|A|^2\right)_x\right] dx + \text{c.c.} \\ &\leq 2|\gamma + i\delta| \int_0^1 [(2|A||B_x| + |B||A_x|)|B||A_{xxx}| \\ &\qquad + (2|B||A_x| + |A||B_x|)|A||B_{xxx}|]\, dx. \end{aligned}$$

But, according to the mean value theorem, if $t > 0$, then $|A(x,t)|^2 \leq 2\phi_0(t) + \phi_1(t)$, and $|B(x,t)|^2 \leq 2\phi_0(t) + \phi_1(t)$. Then, by applying Hölder's inequality we readily obtain $F_3 + F_4 \leq C[\phi_0(t) + \phi_1(t)]\phi_1(t)^{1/2}\phi_3(t)^{1/2}$ for all $t > 0$, where the constant C is as defined at the beginning of the proof (in fact, it depends only on β, γ, and δ). Then, (3.9) yields $d\phi_2/dt = -2\phi_3 + 2\mu\phi_2 + C(\phi_0 + \phi_1)\phi_1^{1/2}\phi_3^{1/2}$ if $t > 0$, and by the argument leading to (3.8) above,

$$\frac{d\phi_2}{dt} \leq 2\mu\phi_2 + \frac{C^2(\phi_0 + \phi_1)^2\phi_1}{4} \quad \text{if } t > 0. \tag{3.10}$$

Now, if (3.8) is multiplied by $\lambda = 1 + 2|\mu| > 0$, the resulting equation is added to (3.10) and the first estimate (3.5) for ϕ_0 and ϕ_1 is used, then the following inequality is obtained: $d(\phi_2 + \lambda\phi_1)/dt \leq -(\phi_2 + \lambda\phi_1) + K$ if $t > 0$, where the constant K is as defined at the beginning of the proof. By using this inequality, the first estimate (3.5) readily follows for ϕ_2. Again, the second estimate (3.5) is obtained similarly, when taking into account that it holds for ϕ_0 and ϕ_1, and using the appropriate linear combination of (3.8) and (3.10). Thus, the proof is complete. □

Now, we show that if ε is sufficiently small, then the solutions of (1.1)–(1.4) satisfy approximately (1.20)–(1.23), in an appropriate uniform sense.

THEOREM 3.4. *If $\gamma < 1$ and $0 < \varepsilon < 1$, let (A, B) be a classical solution of* (1.1)–(1.3) *in $-\infty < x < \infty$, $t \geq 0$, and let the functions V and W be defined by*

$$V(y,t) \equiv \varepsilon^{-4/5} \int_t^{t+\varepsilon^{4/5}} A(y+\tau/\varepsilon, \tau)\, d\tau, \quad W(z,t) \equiv \varepsilon^{-4/5} \int_t^{t+\varepsilon^{4/5}} B(z-\tau/\varepsilon, \tau)\, d\tau.$$

Then, for all $y, z \in \mathbb{R}$, and $t \geq 0$, V and W satisfy

$$V_t = (1+i\alpha)V_{yy} + \mu V - (1+i\beta)V|V|^2 + (\gamma + i\delta)V\langle|W|^2\rangle + \varphi_1(y,t), \tag{3.11}$$

$$W_t = (1+i\alpha)W_{zz} + \mu W - (1+i\beta)W|W|^2 + (\gamma + i\delta)W\langle|V|^2\rangle + \varphi_2(z,t), \tag{3.12}$$

$$V(y+1,t) = V(y,t)\exp(-ia), \quad W(z+1,t) = W(z,t)\exp(ia), \tag{3.13}$$

$$V(y,t) = A(y+t/\varepsilon, t) + \psi_1(y,t), \quad W(z,t) = B(z-t/\varepsilon, t) + \psi_2(z,t), \tag{3.14}$$

where the spatial average $\langle\cdot\rangle$ is defined by (1.24) *and the functions φ_j and ψ_j satisfy, for $j = 1$ and 2,*

$$\|\varphi_j(\cdot,t)\|_\infty, \quad \|\psi_j(\cdot,t)\|_\infty \leq K_0\varepsilon^{1/5} \quad \textit{if } t \geq 0, \tag{3.15}$$

$$\|\varphi_j(\cdot,t)\|_\infty, \quad \|\psi_j(\cdot,t)\|_\infty \leq C_0\varepsilon^{1/5} \quad \textit{if } t \geq T \tag{3.16}$$

for some constants K_0, T, and C_0 such that K_0 and T (respectively, C_0) depend only on $\|A(\cdot,0)\|_{H^2(]0,1[)}$, $\|B(\cdot,0)\|_{H^2(]0,1[)}$, β, γ, δ, and μ (respectively, on β, γ, δ, and μ). Here $H^2(]0,1[)$ is the usual (complexified) Sobolev space and $\|\cdot\|_\infty$ is the sup *norm (if $g \in C^2(\mathbb{R})$, then $\|f\|_{H^2(]0,1[)} = [\int_0^1 (|f(\xi)|^2 + |f'(\xi)|^2 + |f''(\xi)|^2)\, d\xi]^{1/2}$ and $\|g\|_\infty = \sup\{|g(\xi)| : \xi \in \mathbb{R}\}$).*

Proof. We shall only obtain the estimate (3.15) (the proof below extends straightforwardly to obtain (3.16), when using the second estimate (3.5) of Lemma 3.3). As in the proof of Lemma 3.3, every constant depending only on $\|A(\cdot,0)\|_{H^2(]0,1[)}$, $\|B(\cdot,0)\|_{H^2(]0,1[)}$, β, γ, δ, and μ will be denoted always as K.

Let the functions v and w be defined by

$$v(y,t) \equiv A(y+t/\varepsilon, t), \quad w(z,t) \equiv B(z-t/\varepsilon, t), \tag{3.17}$$

which are readily seen to satisfy, for all $y, z \in \mathbb{R}$ and all $t \geq 0$,

$$v_t = (1+i\alpha)v_{yy} + \mu v - (1+i\beta)v|v|^2 + (\gamma+i\delta)v|w(y-2t/\varepsilon, t)|^2, \tag{3.18}$$

$$w_t = (1+i\alpha)w_{zz} + \mu w - (1+i\beta)w|w|^2 + (\gamma+i\delta)w|v(z+2t/\varepsilon, t)|^2, \tag{3.19}$$

$$v(y+1,t) = v(y,t)\exp(-ia), \quad w(z+1,t) = w(z,t)\exp(ia). \tag{3.20}$$

Also, according to Lemma 3.5 (see also (3.17) and (3.20)),

$$\int_0^1 (|v|^2 + |v_y|^2 + |v_{yy}|^2)\, dy + \int_0^1 (|w|^2 + |w_z|^2 + |w_{zz}|^2)\, dz \leq K \quad \text{if } t \geq 0,$$

$$\|v(\cdot,t)\|_\infty + \|w(\cdot,t)\|_\infty \le K \quad \text{if } t > 0 \tag{3.21}$$

for some constant K, where the second inequality follows the first one. Then (3.18), (3.19) implies that $\int_0^1 |v_t|^2 dy + \int_0^1 |w_t|^2 dz \le K$ for all $t > 0$, for some constant K, and (see also (3.20)) for all $y_0, z_0 \in \mathbb{R}$, and all $t_0 \ge 0$,

$$\int_{t_0}^{t_0+1} \left[\int_{y_0}^{y_0+1} (|v|^2 + |v_y|^2 + |v_{yy}|^2 + |v_t|^2)\, dy \right] dt \le K,$$
$$\int_{t_0}^{t_0+1} \left[\int_{z_0}^{z_0+1} (|w|^2 + |w_z|^2 + |w_{zz}|^2 + |w_t|^2)\, dz \right] dt \le K$$

for some constant K. Finally, by imbedding theorems (the Sobolev space $H^{2,1}$ ($]a,b[\times]t_0, t_0+1[$) is continuously imbedded into the Hölder space $C^{1/2,1/4}$ ($[a,b] \times [t_0, t_0+1]$); see, e.g., [11, II-3]) we have

$$|v(y,t_2) - v(y,t_1)| + |w(z,t_2) - w(z,t_1)| \le K(t_2 - t_1)^{1/4} \tag{3.22}$$

if $y, z \in \mathbb{R}$, $0 \le t_1 \le t_2 \le t_1 + 1$, and this estimate and (3.21) are sufficient to obtain the required result.

To see that, first notice that since

$$V(y,t) = \varepsilon^{-4/5} \int_t^{t+\varepsilon^{4/5}} v(y,\tau) d\tau, \qquad W(z,t) = \varepsilon^{-4/5} \int_t^{t+\varepsilon^{4/5}} w(z,\tau) d\tau, \tag{3.23}$$

(3.20) readily implies that (3.13) holds. Also, if $0 < \varepsilon < 1$, then by using (3.22) we obtain, for all $y \in \mathbb{R}$ and all $t \ge 0$,

$$\begin{aligned} |\psi_1(y,t)| &= \varepsilon^{-4/5} \left| \int_t^{t+\varepsilon^{4/5}} (v(y,\tau) - v(y,t)) dt \right| \\ &\le K\varepsilon^{-4/5} \int_t^{t+\varepsilon^{4/5}} (t-\tau)^{1/4} = (4K/3)\varepsilon^{1/5}, \end{aligned} \tag{3.24}$$

and the estimate (3.15) holds for ψ_1 (similarly it is seen that this estimate holds for ψ_2). Finally, to obtain the estimate (3.15) for φ_1 (it is obtained for φ_2 a similar way) observe that φ_1 is given by

$$\varphi_1(y,t) = (1+i\beta)\varepsilon^{-4/5} F_1 + (\gamma + i\delta)\varepsilon^{-4/5}(F_2 + F_3 + F_4), \tag{3.25}$$

as obtained when (3.18) is integrated in $]t, t+\varepsilon^{4/5}[$, and the resulting equation is multiplied by $\varepsilon^{-4/5}$, where

$$F_1 \equiv \int_t^{t+\varepsilon^{4/5}} v(y,\tau)|v(y,\tau)|^2 d\tau - \varepsilon^{4/5} V(y,t)|V(y,t)|^2,$$

$$F_2 \equiv \int_t^{t+\varepsilon^{4/5}} [V(y,t) - v(y,\tau)]\, |w(y - 2\tau/\varepsilon, \tau)|^2 d\tau,$$

$$F_3 \equiv V(y,t) \int_t^{t+\varepsilon^{4/5}} \left[|w(y-2\tau/\varepsilon,\tau)|^2 - |w(y-2\tau/\varepsilon,t)|^2 \right] d\tau,$$

$$F_4 \equiv V(y,t) \left[\varepsilon^{4/5} \langle |W(\cdot,t)|^2 \rangle - \int_t^{t+\varepsilon^{4/5}} |w(y-2\tau/\varepsilon,t)|^2 d\tau \right].$$

Then, to end up the proof, we only need to show that

$$|F_j| \leq K\varepsilon \quad \text{if} \quad t \geq 0 \quad \text{for } j = 1, \dots, 4. \tag{3.26}$$

But the first three estimates (3.26) are readily obtained, by the argument leading to (3.24), when using (3.21) and (3.22), while the fourth one comes from the following expression, which is obtained by means of (3.20):

$$\begin{aligned}
&\int_t^{t+\varepsilon^{4/5}} |w(y - 2\tau/\varepsilon, t)|^2 d\tau \\
&\quad = (\varepsilon/2) \int_{y-2t/\varepsilon}^{y-2t/\varepsilon - 2\varepsilon^{-1/5}} |w(\xi, y)|^2 d\xi \\
&\quad = (\varepsilon/2) \mathrm{Int}(2\varepsilon^{-1/5}) \langle |W(\cdot, t)|^2 \rangle - (\varepsilon/2) \int_0^{\mathrm{Fract}(2\varepsilon^{-1/5})} |w(\xi, t)|^2 d\xi,
\end{aligned}$$

where Int and Fract stand for the integral and fractionary parts, respectively. When (3.26) is substituted into (3.25), the estimate (3.15) is obained for φ_1. Thus, the proof is complete. □

Finally, if ε is sufficiently small, then the solutions of (1.1)–(1.4) that satisfy one of the additional conditions, (1.6) or (1.7), are approximate solutions of (1.25)–(1.27), in an appropriate uniform sense.

THEOREM 3.5. *Under the assumptions of Theorem* 3.4, *let us assume that* (A, B) *satisfy one of the additional assumptions,* (1.6) *or* (1.7). *Then the function* V *is such that, for all* $y \in \mathbb{R}$ *and all* $t \geq 0$,

$$V_t = (1 + i\alpha)V_{yy} + \mu V - (1 + i\beta)V|V|^2 + (\gamma + i\delta)V\langle |V|^2 \rangle + \varphi(y, t), \tag{3.27}$$

$$V(y + 1, t) = V(y, t) \exp(-ia), \tag{3.28}$$

$$V(y, t) = A(y + t/\varepsilon, t) + \psi(y, t), \tag{3.29}$$

where the functions φ *and* ψ *satisfy*

$$\|\varphi(\cdot, t)\|_\infty, \quad \|\psi(\cdot, t)\|_\infty \leq K_0 \varepsilon^{1/5} \quad \text{if } t \geq 0, \tag{3.30}$$

$$\|\varphi(\cdot, t)\|_\infty, \quad \|\psi(\cdot, t)\|_\infty \leq C_0 \varepsilon^{1/5} \quad \text{if } t \geq T \tag{3.31}$$

for some constants K_0, T, *and* C_0 *such that* K_0 *and* T (*respectively,* C_0) *depend only on* $\|A(\cdot, 0)\|_{H^2(]0,1[)}$, β, γ, δ, *and* μ (*respectively, on* β, γ, δ, *and* μ). *Here the average* $\langle \cdot \rangle$, *the norm* $\| \cdot \|_\infty$ *and the space* $H^2(]0, 1[)$ *are as defined in Theorem* 3.4.

Proof. The result is obtained from that in Theorem 3.4 when taking into account that if (A, B) satisfy one of the additional assumptions, (1.6) or (1.7), then $\langle |V|^2 \rangle = \langle |W|^2 \rangle$ for all $t \geq 0$. □

Remark 3.6. According to Theorem 3.4 (respectively, Theorem 3.5), each solution of (1.1)–(1.3) (respectively, of (1.1)–(1.3), (1.6) or (1.1)–(1.3), (1.7)) satisfies approximately, in the uniform sense of the estimates (3.15) (respectively, (3.30)), the model (1.20)–(1.23) (respectively, (1.25)–(1.26)). The estimates (3.16) (respectively, (3.31)) are independent of the particular solution of (1.1)–(1.3) (respectively, of (1.1)–(1.3), (1.6) or (1.1)–(1.3), (1.7)) that is considered, but they apply only for sufficiently large time.

3.3. Some basic properties of the approximate models. Here we give two basic properties of the approximate models, namely, that they are well posed and possess a globally attracting set.

THEOREM 3.7. *If* $\gamma < 1$ *and the functions* A_0 *and* B_0 (*respectively,* A_0) *belong to the complexified space of* $C^{2+\alpha}([0,1])$, *for some* $\alpha \in]0,1[$, *and satisfy* (1.22) (*respectively,* (1.26)), *then the problem* (1.20)–(1.23) (*respectively,* (1.25)–(1.26)) *possesses a unique solution,* $A, B \in$ (*respectively,* $A \in$)$C^{2+\alpha,1+\alpha/2}([0,1] \times [0,T])$ *for all* $T > 0$. *In addition there is a constant* C, *depending only on* α, β, γ, δ, *and* μ, *such that, for each solution of* (1.20)–(1.23) (*respectively,* (1.25)–(1.26)) *there is a constant* T *that satisfies, for* $k = 0$ *and* 1,

$$\|A(\cdot,t)_{C^k([0,1])} + \|B(\cdot,t)\|_{C^k([0,1])} \leq C \quad \left(\textit{respectively, } \|A\|_{C^k([0,1])} \leq C\right)$$

for all $t \geq T$.

Proof. The proof of the first statement is completely similar to that of Theorem 3.1, and the second statement is proven by the ideas in the proof of Lemma 3.3. □

Remark 3.8. If, in addition to the assumptions in Theorem 3.7, $\mu \leq 0$, then the constant C may be taken arbitrarily small; this means that the trivial solutions of (1.20)–(1.23) and (1.25)–(1.26), $A \equiv B \equiv 0$ and $A \equiv 0$, are globally, asymptotically stable in this case. Therefore, if $\mu \leq 0$, then the dynamics of the approximate models are trivial. If $\gamma = 1$ (respectively, $\gamma < 1$), then both approximate models possess solutions that are unbounded as $t \to \infty$ (respectively, that blow up in a finite time), as is seen by considering spatially uniform solutions of both models.

4. Concluding remarks. We have considered the models (1.1)–(1.3), (1.1)–(1.3), (1.6), and (1.1)–(1.3), (1.7) that, as it was explained in §1, are normal forms that apply at the onset of the so-called oscillatory instability in a large variety of physical problems. In the limit $\varepsilon \to 0$, we obtained (formally in §2, and rigorously in §3) two simpler approximate models, (1.20)–(1.23) and (1.25)–(1.26). In addition, we have seen that both models are well posed, and possess a globally attracting set. Some remarks concerning these models are in order.

(A) Both aproximate models possess an inertial manifold of finite dimension. This result may be seemingly proved by extending the analysis in [18], on the standard cubic complex Ginzburg–Landau equation.

(B) The approximate model (1.20)–(1.23) possesses a two-parameter family of travelling waves of the form $V = R\exp(i\omega t + iky + id)$, $W = 0$, or $V = 0$, $W = R\exp(i\omega t + iky + id)$ for appropriate values of the real constants R, ω, k, and d, and a four-parameter family of quasi-periodic waves of the form $V = R_1 \exp(i\omega_1 t + ik_1 y + id_1)$, $W = R_2 \exp(i\omega_2 t + ik_2 y + id_2)$, for appropriate values of the real constants R_1, R_2, ω_1, ω_2, k_1, k_2, d_1, and d_2 (see [19]). These wave-like solutions may lose their stability either under uniform or under nonuniform perturbations; in the second case, the solutions are said to be modulationally unstable. At the onset of the modulational instability, the model (1.20)–(1.23) may be reduced to the Kuramoto–Sivashinsky equation (see [19]) that, as is well known, exhibits chaotic behavior. Therefore, we may expect chaotic solutions of (1.20)–(1.23) for appropriate values of the parameters. The same conjecture can be made in connection with the approximate model (1.25), (1.26).

(C) The particular wave-like solutions of (1.20)–(1.23) and (1.25), (1.26) mentioned in (B) are such that the moduli of V and W are spatially uniform. More general solutions, with spatially nonuniform modulus, are of great physical interest;

in particular, they provide a description of the so-called *blinking states* of the original model (1.8), which have been detected in experiments (see [10], [11]). Both models are expected to possess such solutions. In particular, if $\alpha = \beta = \delta = 0$, then it may be seen that every stationary solution of (1.25), (1.26) is such that either (i) θ =constant or (ii) $R \neq 0$ for all $y \in [0,1]$, where R and θ are the modulus and the argument of V. The stationary solutions of type (i) may be obviously obtained in closed form, in terms of elliptic functions. The solutions of type (ii) are readily seen to be such that $d\theta/dx = c/R^2$ for some constant $c \neq 0$; then R satisfies $R'' + \mu R - R^3 - c^2/R^3 + \gamma R \int_0^1 R^2 dy = 0$, and this equation is again solved in terms of elliptic functions. In addition, the evolution problem (1.25), (1.26) is gradient-like (admits the Lyapunov function $\varphi : H^2(]0,1[) \to \mathbb{R}$ defined by $\varphi(V) \equiv \int_0^1 (|V'(y)|^2 - \mu|V(y)|^2) + \int_0^1 (|V(y)|^4/2)\, dy - \gamma[\int_0^1 |V(y)|^2 dy]^2)$, and the stability of these solutions is readily analyzed. We do not include a complete study of these solutions and of their stability because it has been announced in [12]. Sufficient conditions for the existence of stable solutions of (1.25), (1.26) in the general (nonreal coefficients) case would be of great interest.

REFERENCES

[1] P. COULLET, S. FAUVE, AND E. TIRAPEGUI, *Large scale instability of nonlinear standing waves*, J. Physique Lett., 46 (1985), pp. L787–L791.

[2] S. FAUVE, *Large scale instabilities of cellular flows*, in Instabilities and Nonequilibrium Structures, E. Tirapegui and D. Villarroel, eds., Reidel, Dordrecht, the Netherlands, 1987, pp. 63–88.

[3] P. C. HOHENBERG AND M. C. CROSS, *An introduction to pattern formation in non-equilibrium systems*, in Lecture Notes in Phys., Vol. 268, Springer-Verlag, Berlin, 1987, pp. 55–92.

[4] A. C. NEWELL, *Dynamics of patterns: a survey*, NATO Workshop on Propagation in Nonequilibrium Systems, held at les Houches 1989, Springer-Verlag, Berlin.

[5] M. C. CROSS, *Traveling and standing waves in binary-fluid convection in finite geometries*, Phys. Rev. Lett., 57 (1986), pp. 2935–2938.

[6] ———, *Structure of non-linear traveling-wave states in finite geometries*, Phys. Rev. A, 38 (1988), pp. 3593–3600.

[7] C. M. SURKO AND P. KOLODNER, *Oscillatory traveling-wave convection in a finite container*, Phys. Lett., 58 (1987), pp. 2055–2058.

[8] E. MOSES, J. FINEBERG, AND V. STEINBERG, *Multistability and confined travelling-wave patterns in a convecting binary mixture*, Phys. Rev. A, 35 (1987), pp. 2757–2760.

[9] R. HEINRICHS, G. AHLERS, AND D. S. CANNEL, *Traveling waves and spatial variation in the convection of a binary mixture*, Phys. Rev. A, 35 (1987), pp. 2761–2764.

[10] J. FINEBERG, E. MOSES, AND V. STEINBERG, *Spatially and temporally modulated traveling-wave pattern in convecting binary mixtures*, Phys. Rev. Lett., 61 (1988), pp. 838–841.

[11] P. KOLODNER AND C. M. SURKO, *Weakly non-linear travelling-wave convection*, Phys. Rev. Lett., 61 (1988), pp. 842–845.

[12] E. KNOBLOCH AND J. DE LUCA, *Amplitude equations of travelling wave convection*, preprint, 1990.

[13] C. ALVAREZ-PEREIRA AND J. M. VEGA, *On the pulsating instability of two-dimensional flames*, European J. Appl. Math., 3 (1992), pp. 55–73.

[14] J. M. GHIDAGLIA, *Etude d'écoulements de fluides visqueux incompressibles: comportement por les grands temps et applications aux attracteurs,* Thèse de 3e Cycle, Université Paris Sud, Orsay, 1984.

[15] R. TEMAM, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Springer-Verlag, Berlin, 1988.

[16] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, Berlin, 1981.

[17] O. A. LADYŽENSKAJA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[18] C. R. Doering, J. D. Gibbon, D. D. Holm, and B. Nicolaenko, *Finite Dimensionality in the Complex Ginzburg–Landau Equation*, Contemporary Mathematics, 99, American Mathematical Society, Providence, RI, 1989.

[19] C. Martel and J. M. Vega, *The oscillatory instability of plane wave-fronts of reaction-diffusion systems*, in preparation.

SIAM J. MATH. ANAL.
Vol. 24, No 3, pp. 618-633, May 1993

005

SOME EXISTENCE THEOREMS FOR AN N-DIMENSIONAL PARABOLIC EQUATION WITH A DISCONTINUOUS SOURCE TERM*

ROBERTO GIANNI† AND PAOLA MANNUCCI†

Abstract. The authors consider an n-dimensional semilinear equation of parabolic type with a discontinuous source term arising from combustion theory. The authors prove a local existence for a classical solution having a "regular" free boundary. In this regard, the free boundary is a surface through which the discontinuous source term exhibits a switch-like behaviour. The authors specify conditions under which this solution and its free boundary are global in time. The authors also prove uniqueness and continuous dependence theorems.

Key words. combustion, semilinear parabolic equation, regularity of the solution and of the free boundary, discontinuous source term, uniqueness, continuous dependence

AMS subject classifications. 35K57, 35R35, 80A25

Introduction. In this paper we will study a free boundary problem arising from combustion theory. The equation ruling this phenomenon is parabolic with a discontinuous source term:

$$u_t - \Delta u = \mathbb{H}(u-1) \quad \text{in } Q \times [0, T], \tag{0.1}$$

where $\mathbb{H}$ is the Heaviside function and Q is an open subdomain of $\mathbb{R}^N$.

These kinds of problems were investigated by Norbury and Stuart in [1], where the equation was derived from a combustion problem in a porous medium. In further papers Norbury and Stuart have studied some mathematical aspects of such problems but only for the one-dimensional case. (See [2], [3], [4].)

In this problem the free boundary is the level set $\{u=1\}$ that divides Q into two subdomains Q_1 and Q_2 in which $u>1$ and $u<1$, respectively. We will investigate the regularity of the free boundary, its global in time existence, and the possibility of describing it as a surface of the form $x_i = f(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N, t)$.

In [5] the authors prove analogous results in the one-dimensional case.

For the sake of simplicity, in this paper we will work in two dimensions with $Q = \{(x, y)/0<x<1; 0<y<1\}$, but the techniques we use can be applied to a general domain in $\mathbb{R}^N$ which we can assume to have its boundary S in the class O^2 in the sense specified in p. 9 of [6].

Moreover, we will consider the simple equation (0.1), while in [5] a quite general quasilinear equation was considered. These assumptions are made only to make the proofs simpler and are not going to restrict in any way the validity of the theorems. For example, instead of (0.1) we can consider the equation

$$u_t - \nabla \cdot (K(u)\nabla u) = f(u)\mathbb{H}(u-1),$$

where, as far as the global results are concerned, $K(u)$ is a constant and $f(u)$ is assumed to be a positive nondecreasing function belonging to $C^1(\mathbb{R})$.

More precisely in this paper we will prove a local in time existence theorem for a solution of (0.1) having a "regular" free boundary, then we will find particular boundary conditions and initial data which assure "global" existence of such a solution.

* Received by the editors July 24, 1992; accepted for publication (in revised form) October 5, 1992.

† Dipartimento di Matematica "Ulisse Dini," Università degli Studi di Firenze, Viale Morgagni 67/A, 50134 Firenze, Italy.

Moreover, we will investigate uniqueness and continuous dependence results for problem (0.1).

In the following we will denote with α, $\alpha \in (0, 1)$, any suitable Hölder exponent not specified further.

Where not differently specified, we will use the notation of [6].

1. Local existence. Let us now consider the following two-dimensional parabolic problem:

$$u_t = \Delta u + \mathbb{H}(u-1), \tag{1.1}$$

$$u(x, y, 0) = h(x, y), \tag{1.2}$$

$$u(0, y, t) = K_1(y, t), \tag{1.3a}$$

$$u(1, y, t) = K_2(y, t), \tag{1.3b}$$

$$u(x, 0, t) = K_3(x, t), \tag{1.3c}$$

$$u(x, 1, t) = K_4(x, t), \tag{1.3d}$$

where $(x, y) \in Q = (0, 1) \times (0, 1)$, $t \in (0, T)$. Let $Q_T = Q \times (0, T)$, $\partial_p Q_T$ the parabolic boundary of Q_T, and S_T the lateral boundary of Q_T.

In (1.1), $\mathbb{H}(\eta)$ is the Heaviside function so defined: $\mathbb{H}(\eta) = 1$ if $\eta > 0$ and $\mathbb{H}(\eta) = 0$ if $\eta \leqq 0$. Let us list the assumptions we will use

$$\begin{cases} K_i \in H^{\beta,\beta/2}([0, 1] \times [0, T]), & i = 1, \dots, 4, \\ K_{iy}(y, t) \in H^{\beta,\beta/2}([0, 1] \times [0, T]), & i = 1, 2, \\ K_{ix}(x, t) \in H^{\beta,\beta/2}([0, 1] \times [0, T]), & i = 3, 4, \quad \beta > \frac{3}{4}, \\ K_{it}(x; y, t) \in L_\infty([0, 1] \times [0, T]), & i = 1, \dots, 4, \end{cases} \tag{1.4}$$

$$h(x, y) \in H^{1+\beta}(\bar{Q}), \beta > \tfrac{3}{4}, \tag{1.5}$$

$$|\Delta h| \leqq M, \tag{1.6}$$

$$\begin{cases} A_0 = \{(x, y) \colon h(x, y) = 1\} \text{ admits the representation } y = f_0(x), \quad 0 \leqq x \leqq 1, \\ f_0(x) \in H^{1+\beta}([0, 1]), \\ h > 1 \quad \text{in } \{(x, y) \in \bar{Q} \colon y > f_0(x)\}, \\ h < 1 \quad \text{in } \{(x, y) \in \bar{Q} \colon y < f_0(x)\}, \end{cases} \tag{1.7}$$

(1.8) $h_y(x, y) \geqq \gamma > 0$ in a neighbourhood relative to $\bar{Q}$ of the curve $y = f_0(x)$,

$$\begin{cases} \text{zero order compatibility conditions are satisfied on } \partial_p Q_T, \\ \text{this means that the boundary and initial data} \\ \text{are a continuous function on } \partial_p Q_T. \end{cases} \tag{1.9}$$

Remark 1.1. As far as we are only concerned with a local existence theorem, we can relax the assumptions (1.4)–(1.6). In fact, we can just assume that $K_3(x, t)$, $K_4(x, t)$ are Hölder continuous with exponent $\alpha \in (0, 1)$ and $K_1(y, t)$, $K_2(y, t)$ satisfy (1.4) only in a neighbourhood of $(0, f_0(0), 0)$ and $(1, f_0(1), 0)$ in the plane $\{x = 0\}$ and $\{x = 1\}$, respectively, while $h(x, y)$ satisfies (1.5), (1.6) only in a neighbourhood of the curve $y = f_0(x)$ which is still assumed to belong to $H^{1+\beta}([0, 1])$. This means that the assumptions on the regularity of the boundary and initial data are made only "near" the curve $y = f_0(x)$ while the regularity required for the boundary conditions away from this curve will be the weakest needed to have the existence of a weak solution of the problem, for example, in $W_2^{1,0}(Q_T) \cap C^0(\bar{Q}_T)$.

THEOREM 1.1. *Under the assumptions* (1.4)-(1.9), *a function* $u(x, y, t)$ *exists such that*

$$u\in H^{\alpha,\alpha/2}(\overline{Q_T}),\quad \nabla u\in H^{\alpha,\alpha/2}(\bar{Q}'_T)\quad \forall\varepsilon'>0, \tag{1.10}$$

with $Q'=(0,1)\times(\varepsilon', 1-\varepsilon')$,

$$u_{xx}, u_{yy}, u_{xy}, u_t\in L_2(Q'_T)\quad \forall\varepsilon'>0 \tag{1.11}$$

satisfying (1.2)-(1.3d) *and* (1.1) *almost everywhere. Moreover, for a sufficiently small* T_1 *we have*

$$|u_t|\leqq M\quad \textit{in } Q_{T_1}, \tag{1.12}$$

and denoting with A_{T_1} *the level set*

$$A_{T_1}=\{(x, y, t)\,|\,(x, y)\in\bar{Q}, t\in(0, T_1), u(x, y, t)=1\},$$

we have that A_{T_1} *is the graph of a function of the form* $y=f(x, t)$ *with* $f_x(x, t)\in H^{\alpha,\alpha/2}([0,1]\times[0,T_1])$, $|f_t(x,t)|\leqq M$ *in* $[0,1]\times[0,T_1]$.

The proof of Theorem 1.1 will follow as a consequence of Propositions 1.1, 1.2, and 1.3 below. Let us consider a sequence of approximating problems:

$$u_{nt}=\Delta u_n+\mathbb{H}_n(u_n-1), \tag{1.13}$$

$$u_n(x, y, 0)=h_n(x, y), \tag{1.14}$$

$$u_n(0, y, t)=K_{1n}(0, y, t), \tag{1.15a}$$

$$u_n(1, y, t)=K_{2n}(1, y, t), \tag{1.15b}$$

$$u_n(x, 0, t)=K_{3n}(x, 0, t), \tag{1.15c}$$

$$u_n(x, 1, t)=K_{4n}(x, 1, t), \tag{1.15d}$$

where:

$$\left\{\begin{array}{l} h_n\in C^\infty(\bar{Q}),\ K_{in}\in C^\infty([0,1]\times[0,T]) \\ \text{and they converge, respectively, to } h \text{ and } K_i \\ \text{in the } C^1\text{-norm}, \\ i=1,\dots,4, \end{array}\right. \tag{1.16}$$

$$\left\{\begin{array}{l} \nabla h_n \text{ has norm in } H^\beta(\bar{Q}) \text{ bounded independently of } n \\ \text{and } \Delta h_n \text{ is bounded independently} \\ \text{of } n \text{ in } Q, \end{array}\right. \tag{1.17a}$$

$$\left\{\begin{array}{l} K_{iny}(y,t),\ i=1,2 \text{ and } K_{inx}(x,t),\ i=3,4 \text{ have the norms in } H^{\beta,\beta/2} \text{ bounded} \\ \text{independently of } n \text{ with } \beta \text{ as in (1.4)-(1.7)}, \\ K_{int}(x,t),\ i=1,\dots,4, \text{ have their norm in } L_\infty \text{ bounded independently of } n, \end{array}\right. \tag{1.17b}$$

$$\left\{\begin{array}{l} \mathbb{H}_n \text{ are monotonically increasing functions of their argument}, \\ \mathbb{H}_n\in C^\infty(\mathbb{R}), \\ \lim_n \|\mathbb{H}_n-\mathbb{H}\|_{C^2}=0 \text{ in } \mathbb{R}\backslash(-\varepsilon,\varepsilon)\quad \forall\varepsilon>0, \\ \lim_n \|\mathbb{H}_n-\mathbb{H}\|_{2,\mathbb{R}}=0. \end{array}\right. \tag{1.18}$$

$$\left\{\begin{array}{l} \text{Moreover, we can assume that the boundary data (1.14)-(1.15d)} \\ \text{satisfy the zero-order compatibility conditions.} \end{array}\right. \tag{1.19}$$

From (1.16), (1.17a), it follows that, for sufficiently large n, the level set $A_{0n} = \{(x, y): h_n(x, y) = 1\}$ can be described as the graph of a function of the form $y = f_{0n}(x)$, $0 \leqq x \leqq 1$, and $f_{0n}(x) \in H^{1+\beta}([0, 1])$, f_{0n} bounded independently of n in $H^{1+\beta}$. Here and in the following M will denote any constant independent of n. Let $Q' = (0, 1) \times (\varepsilon', 1 - \varepsilon')$, for all $\varepsilon' > 0$.

PROPOSITION 1.1. *Under the assumptions* (1.16)-(1.19), *problem* (1.13)-(1.15d) *admits a unique classical solution* $u_n(x, y, t)$ *such that*

$$u_n(x, y, t) \in H^{2+\alpha, 1+\alpha/2}(Q_T), \tag{1.20}$$

and the following estimates hold:

$$\max_{Q_T} |u_n| \leqq M, \tag{1.21}$$

$$|u_n|_{Q_T}^{(\alpha)} \leqq M, \tag{1.22}$$

$$|u_n|_{Q'_T}^{(1+\alpha)} \leqq M(\varepsilon') \quad \forall \varepsilon' > 0, \tag{1.23}$$

$$\|u_{nxx}, u_{nyy}, u_{nxy}, u_{nt}\|_{2, Q'_T} \leqq M(\varepsilon') \quad \forall \varepsilon' > 0 \tag{1.24}$$

with M *and* $M(\varepsilon')$ *independent of* n.

Proof. By means of Theorem 6.4, p. 460 and Theorem 12.1, p. 223 of [6] we obtain (1.20). The estimate (1.21) follows from Theorem 2.1 of [6, p. 425].

Theorem 1.1 of [6, p. 419] gives the uniform Hölder continuity with exponent α of the function $u_n(x, y, t)$ and hence estimate (1.22). Estimates (1.23) and (1.24) follow from inequality (10.12) p. 355 and the corollary at the end of Theorem 9.1, p. 341 of [6].

Remark 1.2. Because of the uniform Hölder continuity of $\{u_{ny}\}$ it is obvious that, in the (x, y)-plane, a neighbourhood Q_1 of the curve $y = f_0(x)$ can be chosen such that, in $Q_{1T_0} = Q_1 \times (0, T_0)$ (with T_0 sufficiently small); we have

$$u_{ny}(x, y, t) \geqq \gamma > 0 \quad \forall n > N, \quad \text{with } N \text{ sufficiently large}, \tag{1.25}$$

with γ, Q_1, T_0 independent of n.

PROPOSITION 1.2. *Under the assumptions* (1.16)-(1.19), *we have*

$$|u_{nt}| \leqq M \quad \forall n, \tag{1.26}$$

in Q_{T_1} *with* T_1 *suitably small*; *here* M *does not depend on* n.

Proof. Let us define $\hat{Q}_1 = \{(x, y)/x \in (0, 1); f_0(x) - \varepsilon < y < f_0(x) + \varepsilon\}$ and $\hat{Q}_{1T_1} = \hat{Q}_1 \times (0, T_1)$, with T_1 so small that $u_n < 1 - \varepsilon^*$ on $\{(x, y, t)/x \in (0, 1);\ y = f_0(x) - \varepsilon;\ 0 < t < T_1\}$ and $u_n > 1 + \varepsilon^*$ on $\{(x, y, t)/x \in (0, 1); y = f_0(x) + \varepsilon; 0 < t < T_1\}$ for all n sufficiently large and a suitable ε^*. (This is possible because of (1.22).) Moreover, we can assume $\hat{Q}_{1T_1} \subset Q_{1T_0}$ (see Remark 1.2).

In $\hat{Q}_{1T_1}$, $v_n(x, y, t) = u_{nt}(x, y, t)$ satisfies the following system of equations:

$$\begin{aligned} &v_{nt} - \Delta v_n = \mathbb{H}'_n(u_n - 1) v_n \quad \text{in } \hat{Q}_{1T_1}, \\ &v_n(x, y, t) = D_n(x, y) \quad \text{on } \partial_p \hat{Q}_{1T_1}, \end{aligned} \tag{1.27}$$

D_n is a function bounded independently of n; this result can be obtained using assumptions (1.16), (1.17a), (1.17b) and applying Theorem 10.1 of [6, p. 204] to the problem satisfied by $v_n = u_{nt}$.

This implies that, regarding $\mathbb{H}'_n(u_n - 1) v_n$ as a known source term, we can represent the solution v_n of (1.27) as the sum of two functions v_{1n} and v_{2n}, where v_{1n} is the solution of the homogeneous heat equation with the same initial and boundary data of problem (1.27) and v_{2n} is the solution of equation (1.27) in $\hat{Q}_{1T_1}$ with zero Dirichlet boundary data.

It is quite simple to prove that $|v_{1n}| \leqq M$ in $\hat{Q}_{1T_1}$.

Using the Green function G, we can express v_{2n} in the following form:

$$v_{2n} = \int_0^t \int_{\hat{Q}_1} G(x, y, \xi_1, \xi_2, t, \tau)\mathbb{H}'_n(u_n - 1)v_n \, d\xi_1 \, d\xi_2 \, d\tau. \tag{1.28}$$

Hence we have

$$v_n = v_{1n} + \int_0^t \int_{\hat{Q}_1} G(x, y, \xi_1, \xi_2, t, \tau)\mathbb{H}'_n(u_n - 1)v_n \, d\xi_1 \, d\xi_2 \, d\tau, \tag{1.29}$$

where v_{1n} is regarded as a bounded known term.

Let us define

$$\Omega(t) = \max_{\hat{Q}_1 \times (0,t)} |v_n(x, y, t)|, \qquad N(t) = \max_{\hat{Q}_1 \times [0,t]} |v_{1n}(x, y, t)|.$$

Then we have

$$\Omega(t) \leqq N(t) + \left| \int_0^t \Omega(\tau) \int_{\hat{Q}_1} C(t-\tau)^{-1} \exp\left(\frac{-C(|x-\xi_1|^2 + |y-\xi_2|^2)}{t-\tau} \right) \cdot \mathbb{H}'_n(u_n - 1) \, d\xi_1 \, d\xi_2 \, d\tau \right|.$$

Now, using (1.25), we can multiply and divide $\mathbb{H}'_n(u_n - 1)$ by u_{ny}, thus obtaining

$$\Omega(t) \leqq N(t) + C\left| \int_0^t \Omega(\tau) \int_{\hat{Q}_1} (t-\tau)^{-1} \exp\left(\frac{-C(|x-\xi_1|^2 + |y-\xi_2|^2)}{t-\tau} \right) \cdot \frac{d}{dy}(\mathbb{H}_n(u_n - 1)) \frac{1}{u_{ny}} \, d\xi_1 \, d\xi_2 \, d\tau \right|.$$

Integrating in ξ_2, we get

$$\Omega(t) \leqq N(t) + \frac{C}{\gamma} \left| \int_0^t \Omega(\tau) \int_0^1 (t-\tau)^{-1} \exp\left(\frac{-C|x-\xi_1|^2}{t-\tau} \right) d\xi_1 \, d\tau \right|.$$

Now we multiply and divide the argument of the above integral by $|x-\xi_1|^{1/2}/(t-\tau)^{1/4}$, and, keeping in mind that the quantity

$$\exp\left(\frac{-C|x-\xi_1|^2}{t-\tau} \right) \frac{|x-\xi_1|^{1/2}}{(t-\tau)^{1/4}}$$

is bounded, we obtain (choosing if necessary a new constant C)

$$\Omega(t) \leqq N(t) + \frac{C}{\gamma} \int_0^t \Omega(\tau) \int_0^1 \frac{(t-\tau)^{-3/4}}{|x-\xi_1|^{1/2}} \, d\xi_1 \, d\tau.$$

Now if we set

$$C_0 = \max_{x \in (0,1)} \int_0^1 |x-\xi_1|^{-1/2} \, d\xi_1,$$

we have

$$\Omega(t) \leqq N(t) + \frac{CC_0}{\gamma} \int_0^t \Omega(\tau)(t-\tau)^{-3/4} \, d\tau.$$

Then, multiplying by $CC_0/(t-\tau)^{3/4}\gamma$ both sides of the inequality and integrating in t, we get

$$\begin{aligned}\Omega(t)-N(t)&\leqq\frac{CC_0}{\gamma}\int_0^t N(\tau)(t-\tau)^{-3/4}\,d\tau\\&\quad+\frac{C^2C_0^2}{\gamma^2}\int_0^t(t-\tau)^{-3/4}\int_0^\tau\Omega(\theta)(\tau-\theta)^{-3/4}\,d\theta\,d\tau\\&=\frac{CC_0}{\gamma}\int_0^t N(\tau)(t-\tau)^{-3/4}\,d\tau\\&\quad+\frac{C^2C_0^2}{\gamma^2}\int_0^t\Omega(\theta)\int_\theta^t(t-\tau)^{-3/4}(\tau-\theta)^{-3/4}\,d\tau\,d\theta.\end{aligned}$$

Integrating this last term by means of the substitution $\tau=\theta+(t-\theta)z$, we obtain

$$\Omega(t)-N(t)\leqq\frac{CC_0}{\gamma}\int_0^t N(\tau)(t-\tau)^{-3/4}\,d\tau+\frac{C^2C_0^2}{\gamma^2}\int_0^t M(t-\theta)^{-1/2}\Omega(\theta)\,d\theta,$$

where $M=\int_0^1 z^{-3/4}(1-z)^{-3/4}\,dz$.

Iterating this procedure as in [7, Chap. 3, § 15], we finally obtain for $\Omega(t)$ an integral inequality of the following kind:

$$\Omega(t)\leqq F(t)+\int_0^t K(t-\tau)\Omega(\tau)\,d\tau,$$

with $K(t-\tau)$ bounded independently of n and $F(t)$ depending on $N(t)$ but not on n.

Applying Gronwall's lemma, we get $\Omega(t)\leqq M$ in $(0,T_1)$ with M independent of n.

Since u_{nt} is bounded uniformly on n in $\hat{Q}_{1T_1}$, is straightforward to prove, by means of standard theorems (see [6, Thm. 5.1, p. 444]), that the same result holds true in Q_{T_1}.

PROPOSITION 1.3. *Under the assumptions* (1.16)-(1.19), *in* $Q_{T_1}=Q\times(0,T_1)$, *with* T_1 *suitable small, the level set* $A^n_{T_1}=\{(x,y,t)/(x,y)\in Q;\ t\in(0,T_1)/u_n(x,y,t)=1\}$ *is represented by a regular surface* $y=f_n(x,t)$ *with*

(1.30) $$|f_{nx}|^{(\alpha)}\leqq M\quad in\ [0,1]\times[0,T_1],$$

(1.31) $$|f_{nt}|\leqq M\quad in\ (0,1)\times(0,T_1).$$

Proof. Clearly we have that $A^n_0=\{(x,y)\in Q/h_n(x,y)=1\}$ is a "regular" curve on the plane $t=0$, and if we choose n sufficiently large this set has a distance greater than $\varepsilon/2$ from the curves $y=f_0(x)-\varepsilon$, $y=f_0(x)+\varepsilon$; then for a sufficiently large n and small T_1, $A^n_{T_1}$ lies in $\hat{Q}_{1T_1}$. (Remember in this regard that u_n are uniformly Hölder continuous.) Then, using (1.25), Propositions 1.1 and 1.2, we can apply the inverse function theorem to the function u_n to get that $A^n_{T_1}$ is representable in the form $y=f_n(x,t)$ and estimates (1.30), (1.31) hold.

Let us now turn to the proof of Theorem 1.1.

Proof of Theorem 1.1. We now consider the sequences $\{u_n\}$, $\{f_n\}$ of solutions of the approximating problems (1.13)-(1.15d). Using (1.21)-(1.24), (1.26), (1.30)-(1.31), we see that it is possible to extract from $\{u_n\}$ and from $\{f_n\}$ two subsequences u_{nk} and

f_{nk} that converge, respectively, to the functions u and f, in the following sense:

$$\begin{cases} u_{nk} \xrightarrow{C^0} u \in H^{\alpha,\alpha/2}(\bar{Q}_T) \quad \forall T, \\ \nabla u_{nk} \xrightarrow{C^0} \nabla u \in H^{\alpha,\alpha/2}(\bar{Q}'_T) \quad \forall T, \quad \forall \varepsilon' > 0, \text{ with} \\ Q'_T = (0,1)\times(\varepsilon', 1-\varepsilon')\times(0,T), \\ u_{nkxx};\ u_{nkyy};\ u_{nkxy},\ u_{nkt} \xrightarrow{w-L_2} u_{xx},\ u_{yy},\ u_{xy},\ u_t \in L_2(Q'_T) \quad \forall \varepsilon' > 0, \\ (\text{“}w-L_2\text{” weak convergence in } L_2), \\ |u_t| \leqq M \quad \text{in } Q_{T_1}, \\ f_{nk}(x,t) \xrightarrow{C^0} f(x,t) \text{ with } f(x,t) \text{ Lipschitz continuous in } x \text{ and } t \text{ in} \\ \quad \cdot [0,1]\times[0,T_1], \\ f_{nkx}(x,t) \xrightarrow{C^0} f_x(x,t) \in H^{\alpha,\alpha/2}([0,1]\times[0,T_1]). \end{cases} \tag{1.32}$$

Here T_1 is the constant determined in Propositions 1.2 and 1.3.

To prove that $u(x,y,t)$ is a solution of problem (1.1)-(1.3d) we have to show that

$$\mathbb{H}_n(u_n - 1) \xrightarrow{w-L_2} \mathbb{H}(u-1).$$

We can easily obtain that $\mathbb{H}_n(u_n-1) \xrightarrow{w-L_2} \phi(x,y,t)$, and then

$$u_t - \Delta u = \phi(x,y,t) \quad \text{in } Q_T, \tag{1.33}$$

with $\phi(x,y,t) \in L_\infty(Q_T)$, $\phi(x,y,t) = 0$ if $u(x,y,t) < 1$; $\phi(x,y,t) = 1$ if $u(x,y,t) > 1$; $0 \leqq \phi(x,y,t) \leqq 1$ in $\mathbb{A} = \{(x,y,t) \in Q_T\colon u(x,y,t) = 1\}$.

However, it can be proved that either meas $\mathbb{A} = 0$ or $\phi(x,y,t) = 0$ almost everywhere in $\mathbb{A}$.

In fact, using Lemma A.4 of [8, p. 53], since $\nabla u(x,y,t)$ and $u_t(x,y,t)$ are in L_2, we get $u_t \equiv u_x \equiv u_y \equiv 0$ almost everywhere in $\mathbb{A}$. Still applying Lemma A.4 of [8, p. 53] to the functions $u_x(x,y,\bar{t})$ and $u_y(x,y,\bar{t})$ (remember that because of Fubini's lemma, $u_{xx}(x,y,\bar{t})$ and $u_{yy}(x,y,\bar{t})$ are in $L_2((0,1)\times(0,1))$, for almost all $\bar{t}$), we get $u_{xx}(x,y,\bar{t}) \equiv u_{yy}(x,y,\bar{t}) \equiv 0$ almost everywhere in $\mathbb{A} \cap \{t = \bar{t}\}$, for almost all $\bar{t}$. Hence $u_{xx}(x,y,t) \equiv u_{yy}(x,y,t) \equiv 0$ almost everywhere in $\mathbb{A}$.

Hence, because of (1.33), we have that $\phi(x,y,t) = 0$ almost everywhere in $\mathbb{A}$, then $\phi(x,y,t) \equiv \mathbb{H}(u-1)$ almost everywhere, that is, $u(x,y,t)$ is a solution of (1.1)-(1.3d). To be sure that $y = f(x,t)$ is the representation of the level set $\{u = 1\}$ it only remains to prove the inequalities $u > 1$ in $Q^1_{T_1} = \{(x,y,t) \in Q_{T_1};\ y > f(x,t)\}$, and $u < 1$ in $Q^2_{T_1} = \{Q \backslash \bar{Q}^1_{T_2}\}$. They can be easily obtained by means of the strong maximum principle, so we will not go into the details of the proof.

Proposition 1.4a. *Let $u(x,y,t)$ be the solution of* (1.1)-(1.3d). *Under assumptions* (1.4)-(1.9) *and the assumption $K_{1t} \in H^{\alpha,\alpha/2}([0,1]\times[0,T])$, $i = 1,\dots,4$, then $u_t \in H^{\gamma,\gamma/2}(\mathscr{Q}\times[\tau,T_1])$, $\gamma \in (0,\alpha]$, for all $\tau \in (0,T_1)$, with T_1 suitably small, where $\mathscr{Q}$ is any closed domain contained in $\bar{Q}$ and not containing the corners of the square.*

Proof. Clearly it suffices to work in $\hat{Q}_{1T_1}$ instead of Q_T because in $Q_T \backslash \hat{Q}_{1T_1}$ the result follows from standard theorems on heat equation. Let us consider the solution u_n of the approximating problem (1.13)-(1.15d) under assumptions (1.16)-(1.18) and

$K_{\text{int}} \in H^{\alpha,\alpha/2}$, $|K_{\text{int}}|^{(\alpha)} \leqq M$, for all n. It is important to notice that, using inequality (10.12), p. 155, and the corollary of [6, p. 342], it is easy to prove:

$$|u_{nxx}, u_{nyy}, u_{nxy}, u_{nt}|_q \leqq M, \qquad q = 1/(1-\beta); \tag{A.1}$$

$$|\nabla u_n(x, y, t)|^{(1-4(1-\beta))} \leqq M \tag{A.2}$$

(here β is the same of assumptions (1.4)-(1.9), (1.16)-(1.18)) in $\mathcal{Q} \times [0, T]$. At this point setting $v_n = u_{nt}$, we have that (1.27) can be written in the alternative form:

$$v_{nt} - \Delta v_n + \frac{\mathbb{H}_n(u_n-1)}{u_{ny}} v_{ny} - \frac{\mathbb{H}_n(u_n-1)u_{nyy}}{(u_{ny})^2} v_n = \frac{\partial}{\partial y}\left(\frac{\mathbb{H}_n(u_n-1)v_n}{u_{ny}}\right),$$

where $[\mathbb{H}_n(u_n-1)]/u_{ny}$, $[\mathbb{H}_n(u_n-1)u_{nyy}]/(u_{ny})^2$, $[\mathbb{H}_n(u_n-1)v_n]/u_{ny}$, are regarded as known coefficients belonging respectively to L_∞, L_q, L_∞ (remember for this purpose that from Proposition 1.3, v_n is uniformly bounded in Q_{T_1}). At this point, using Theorem 10.1 of [6, p. 204] we get that

$$|v_n|^{(7)} \leqq M(\tau, \mathcal{Q}) \quad \forall n \text{ on } \mathcal{Q} \times [\tau, T_1] \quad \forall \tau \in (0, T_1), \quad \gamma \in (0, \alpha]. \tag{A.3}$$

Finally, using a diagonal technique, we get Proposition 1.4a.

Remark 1.4b. It is easy to observe that γ and $M(\tau, \mathcal{Q})$ in (A.3) depend on τ, $\mathcal{Q}$ and on the boundary data. This implies that if we consider boundary data whose norms are bounded in their respective spaces by a certain fixed constant $K > 0$, then u_{nt}, u_t are uniformly Hölder continuous with the same Hölder exponent γ. This will turn out to be very useful in the continuous dependence theorem.

In the following theorem we want to prove an important generalization of Theorem 1.1. Namely, we will relax the assumptions (1.5) and (1.6), which guarantee the boundedness of u_t up to the boundary.

THEOREM 1.2. *Under the assumptions* (1.4), (1.7)-(1.9), *if we substitute* (1.5) *with*

$$h(x, y) \in H^{1+\gamma}(\bar{Q}), \qquad \gamma \in (0, 1) \tag{1.5bis}$$

and if we drop (1.6), *we can prove that a function $u(x, y, t)$ exists, such that $u \in H^{\alpha,\alpha/2}(\overline{Q_T})$. For a suitable $\alpha \in (0, 1)$, u_t, u_{xx}, u_{yy}, $u_{xy} \in L_2(Q' \times [\tau, T])$, $\nabla u \in H^{\gamma'}(\bar{Q}'_T)$, $\gamma' \in (0, \gamma]$, for all $\tau > 0$, for all $\varepsilon' > 0$, where $Q' = (0, 1) \times (\varepsilon', 1-\varepsilon')$, $|u_t| \leqq M(\tau)$ in $Q \times [\tau, T_1]$ for all $\tau > 0$ and T_1 suitably small, u satisfies* (1.2)-(1.3d) *and* (1.1) *almost everywhere, and has a level set $\{u = 1\}$ which is representable by a function $y = f(x, t)$ which is Lipschitz continuous with respect to x and t in any subdomain of the form $[0, 1] \times [\tau, T_1]$ and continuous in $[0, 1] \times [0, T_1]$.* (*Equation* (1.5bis) *is different from* (1.5), *where β had to be greater than $\frac{3}{4}$.*)

Proof. Proceeding as we did in Theorem 1.1 we can show that a solution $u(x, y, t)$ of problem (1.1)-(1.3d) exists, having $u \in H^{\alpha,\alpha/2}(\bar{Q}_T)$. As far as the Hölder continuity of ∇u is concerned, it can be easily obtained representing u_n (u_n is, as usual, the solution of a suitable approximating problem) as the sum of two functions u_{1n}, u_{2n}; the first solves the homogeneous heat equation with the boundary data of u_n and the second solves the heat equation with $\mathbb{H}_n(u_n-1)$ as a source term and zero boundary data. At this point the uniform Hölder continuity of ∇u_n in $\bar{Q}'_T$ is obtained in the following way: we split u_{1n} in two functions $\bar{u}_{1n}$ and $\hat{u}_{1n}$. The first one satisfies the heat equation with the same lateral boundary conditions of u_{1n} and "regular" initial data and the second one satisfies the heat equation with zero lateral boundary conditions and suitable initial data. At this point the desired result follows by applying inequality (10.12) of [6, p. 355] and the corollary at the end of Theorem 9.1 of [6, p. 341] to the boundary value problems satisfied by u_{2n}, $\bar{u}_{1n}$; moreover, since $\hat{u}_{1n}$ can be seen as the restriction to the domain $\bar{Q}_T$ of a function $z(x, y, t)$ solving a Cauchy problem for the heat equation in the half plane $\{t \geqq 0\}$ (z_n is obtained from $\hat{u}_{1n}$ by reflection), the estimates on $\nabla \hat{u}_{1n}$ follow from Theorem 11.1 of [6, p. 211]. As far as the L_2-estimates of u_t, u_{xx}, u_{yy}, u_{xy} are concerned, they can be easily obtained by applying inequality (10.12) of [6, p. 355] to the problem solved by u_n. The estimate of Theorem 1.2 on the Lipschitz continuity of $f(x, t)$ follows from an estimate on u_t analogous to (1.26); this can be obtained as done in Proposition 1.2. In fact, in $\hat{Q}_{1T_1}$ we have that u_n can be written as the sum of two functions u_{1n} and u_{2n}, solutions of the following systems of equations:

$$u_{1nt} - \Delta u_{1n} = 0, \tag{1.34}$$

$$u_{1n} = u_n \quad \text{on } \partial_p \hat{Q}_{1T_1}, \tag{1.35}$$

$$u_{2nt} - \Delta u_{2n} = H_n(u_n - 1), \tag{1.36}$$

$$u_{2n} = 0 \quad \text{on } \partial_p \hat{Q}_{1T_1}. \tag{1.37}$$

Splitting u_{1n} in two functions, $\bar{u}_{1n}$ and $\hat{u}_{1n}$, as done before and using the Green's function to represent $\hat{u}_{1n}$, the first estimate of Theorem 16.3 of [6, p. 413] together with assumptions (1.17b) and (1.5bis) yields the following inequality:

$$|u_{1nt}| \leqq M_1(\varepsilon) + \frac{M_2(\varepsilon)}{t^{(1/2)+\varepsilon}} \quad \forall \varepsilon \in (0, \tfrac{1}{2}),$$

with $M_1(\varepsilon)$ and $M_2(\varepsilon)$ independent of n (again we use the fact that $\hat{u}_{1n}$ can be seen as the restriction to the domain $\overline{Q_T}$ of a function $z_n(x, y, t)$ solving a Cauchy problem for the heat equation in the half plane $\{t \geqq 0\}$).

With regard to u_{2n}, differentiating (1.36) with respect to t, we obtain for $u_{2nt} = v_{2n}$,

$$v_{2nt} - \Delta v_{2n} = H_n'(u_n - 1) v_n,$$

$$v_{2n} = 0 \quad \text{on } \partial \hat{Q}_1 \times [0, T_1],$$

$$v_{2n}(x, y, 0) = H_n(h_n - 1),$$

where $v_n = u_{nt}$.

Let $v_{1n} = u_{1nt}$. Then we have

$$\begin{aligned} v_n = v_{1n} + v_{2n} = v_{1n} &+ \int_0^t \int_{\hat{Q}_1} G(x, y, t, \xi_1, \xi_2, \tau) H_n'(u_n - 1) v_n \, d\xi_1 \, d\xi_2 \, d\tau \\ &+ \int_{\hat{Q}_1} G(x, y, t, \xi_1, \xi_2, 0) H_n(h_n - 1) \, d\xi_1 \, d\xi_2. \end{aligned} \tag{1.38}$$

Using again estimates of Theorem 16.3 of [6], we easily get that

$$\left| \int_{\hat{Q}_1} G(x, y, t, \xi_1, \xi_2, 0) H_n(h_n - 1) \, d\xi_1 \, d\xi_2 \right| \leqq \frac{C(\varepsilon)}{t^{(1/2)+\varepsilon}}, \quad \text{with } \varepsilon \in (0, \tfrac{1}{2}).$$

Hence we have

$$|v_n| \leqq M_3(\varepsilon) + \frac{M_4(\varepsilon)}{t^{(1/2)+\varepsilon}} + \int_0^t \int_{\hat{Q}_1} G(x, y, t, \xi_1, \xi_2, \tau) H_n'(u_n - 1) |v_n| \, d\xi_1 \, d\xi_2 \, d\tau,$$

with $M_3(\varepsilon)$ and $M_4(\varepsilon)$ independent of n.

Let $\hat{v}_n(x, y, t)$ be the solution of the following integral equation:

$$\hat{v}_n = M_3(\varepsilon) + \frac{M_4(\varepsilon)}{t^{(1/2)+\varepsilon}} + \int_0^t \int_{\hat{Q}_1} G(x, y, t, \xi_1, \xi_2, \tau) H_n'(u_n - 1) \hat{v}_n \, d\xi_1 \, d\xi_2 \, d\tau.$$

We have that

$$|v_n| \leqq \hat{v}_n.$$

Moreover, following the procedure of [7, Chap. 3, § 15], we obtain

$$|v_n| \leqq \hat{v}_n \leqq C_1(\varepsilon) + \frac{C_2(\varepsilon)}{t^{(1/2)+\varepsilon}}.$$

Now it is straightforward to prove, by means of standard results on parabolic equations, that the same estimate holds in the whole of the domain Q_{T_1}.

The last inequality allows us to get an analogous estimate for f_{nt} and hence for f_t.

The continuity of the function $f(x, t)$ up to $\{t = 0\}$ is an easy consequence of the assumptions and the Hölder continuity of $u(x, y, t)$. This concludes the proof.

In general it seems possible to extend the proof when $\nabla h \in C^0(\bar{Q})$.

Remark 1.3. It is important to stress that assumptions (1.7) can be relaxed. In fact, we can only require that $A_0 = \{(x, y) \in Q / h(x, y) = 1\}$ is represented by a regular parametric curve

$$\text{(1.7bis)} \qquad \begin{cases} x = x(s), \\ y = y(s), \quad 0 \leqq s \leqq 1, \end{cases}$$

with $x(s), y(s) \in H^{1+\beta}([0, 1])$, $x'^2(s) + y'^2(s) \neq 0$ for all $s \in [0, 1]$ and such that ∇h does not vanish at any point of the curve. As a result, in this case we will get that the interface can be represented by a Lipschitz continuous parametric regular surface of the form

$$x = f(r, t), \quad y = g(r, t) \quad (t \text{ as usual represents the time variable})$$

so that a solution of problem (1.1)-(1.3d) will be now represented by the triplet of functions (u, f, g).

The proof is similar to the one above, except Proposition 1.2. In fact, in this case, we have to divide the domain $\hat{Q}_{1T_1}$, in a finite number of subdomains $(\hat{Q}_{1T_1})_i$, $i = 1, \ldots, N$, such that in each $(\hat{Q}_{1T_1})_i$, at least one between u_x and u_y is always different from zero. This is always possible because of the assumptions and of the Hölder continuity of ∇u. At this point an estimate of the integral (1.29) can be done as in Proposition 1.2.

Remark 1.4. It is important to stress that, as far as we are concerned, with a local existence theorem, the domain Q can be chosen with a different shape as long as it is sufficiently smooth; for example, with its boundary in C^2. In fact, proceeding as we did in Proposition 1.1, we can still prove that ∇u_n is uniformly continuous; then if $\nabla h_n \neq 0$ on A_0 for all $n > N$, it remains different from zero in $U \times [0, \hat{T}]$ with U a suitable neighbourhood of A_0 and $\hat{T}$ sufficiently small. At this point the proof goes on as before.

With regard to the global existence theorems which we will prove in the following section, it is easy to realize that the domain Q, which will be chosen to be a square, can be generalized, and we can consider a more general domain of the form $Q = \{(x, y): 0 < x < 1;\ F_1(x) < y < F_2(x)$, with $F_1(x) < F_2(x)\}$. Global existence of the interface will be proved until it touches one between the surfaces $y = F_i(x)$, $i = 1, 2$.

2. Some cases of global existence. From the previous section we see that, for proving the existence of a "regular" free boundary in Q_T, under the assumptions of Theorem 1.1, the crucial point is to find an estimate which bounds away from zero, uniformly in n, $u_{ny}(x, y, t)$ in Q_T (u_n is, as usual, the solution of the approximating problem (1.13)-(1.15d)). (In general, as we have seen in Remark 1.3, it is sufficient to have one between u_{nx} or u_{ny} bounded away from zero in Q_T.) Following the sketch of the proof of Theorem 1.1 it is easy to prove that, if we have an estimate of the type $u_{ny} > 0$ on the parabolic boundary $\partial_p Q_T$, a "regular" free boundary exists in all Q_T. In fact, if we write down an approximating problem as in Theorem 1.1 and we set $u_{ny} = v_n$, we have that v_n solves the parabolic equation

$$v_{nt} - \Delta v_n = \mathbb{H}'_n(u_n - 1)v_n,$$

and $v_n > 0$ on $\partial_p Q_T$. This implies that $v_n \geqq 0$ in $\overline{Q_T}$. Regarding $\mathbb{H}'_n(u_n - 1)v_n$ as a positive source term and applying the maximum principle, we get $v_n = u_{ny} \geqq \delta > 0$ in $\bar{Q}_T$, with δ independent of n. At this point the global existence theorem we are looking for can be proved by means of the same techniques of Theorem 1.1.

We give below some simple examples. More general results can be proved essentially in the same way. Naturally, it is possible that, in this regard, stronger assumptions on the boundary data have to be done.

(a) Dirichlet Problem.

$$u_t = \Delta u + \mathbb{H}(u-1), \tag{2.1}$$

$$u(x, y, 0) = h(x, y), \tag{2.2}$$

$$u(0, y, t) = K_1(y, t), \tag{2.3a}$$

$$u(1, y, t) = K_2(y, t), \tag{2.3b}$$

$$u(x, 0, t) = K_3(x, t), \tag{2.4a}$$

$$u(x, 1, t) = K_4(x, t), \tag{2.4b}$$

with

$$\begin{cases} h_y(x, y) \geqq \gamma > 0 \quad \text{in } Q \\ K_3(x, t) < 1, \quad K_4(x, t) > 1, \quad K_{1y}(y, t) \geqq \gamma > 0, \\ K_{2y}(y, t) \geqq \gamma > 0 \quad \text{in } (0, 1) \times (0, T). \end{cases} \tag{2.5}$$

If (2.5) and all the assumptions of Theorem 1.1 are satisfied, and if K_3 and K_4 satisfy the following conditions

$$K_{3t} - K_{3xx} \leqq 0, \qquad K_{4t} - K_{4xx} \geqq 1, \tag{2.6}$$

since the approximating boundary data can be chosen such that they still satisfy (2.5), (2.6), an estimate on the sign of u_{ny} on the boundary follows immediately from the maximum principle and Hopf's theorem, and hence a global result can be proved as in the previous section.

Remark 2.1. Let us observe that the first assumption of (2.5) can be relaxed, requiring h_y to be nonnegative in Q and strictly positive on the curve $y = f_0(x)$.

(b) Mixed Problem. A special problem with mixed boundary conditions can be studied in a simple way. In fact, if we consider

$$u_t = \Delta u + \mathbb{H}(u-1), \tag{2.1}$$

$$u(x, y, 0) = h(x, y), \tag{2.2}$$

$$u(0, y, t) = K_1(y, t), \tag{2.3a}$$

$$u(1, y, t) = K_2(y, t), \tag{2.3b}$$

$$u_y(x, 0, t) = K_3(x, t), \tag{2.7a}$$

$$u_y(x, 1, t) = K_4(x, t), \tag{2.7b}$$

under the assumptions (2.5) for $h(x, y)$, $K_1(y, t)$, $K_2(y, t)$, $K_3(x, t) \geqq 0$, and $K_4(x, t) \geqq 0$, the proof follows analogously to the one sketched in the previous case (a).

(c) Neumann Problem.

$$u_t = \Delta u + \mathbb{H}(u-1), \tag{2.1}$$

$$u(x, y, 0) = h(x, y), \tag{2.2}$$

$$u_x(0, y, t) = K_1(y, t), \tag{2.8a}$$

$$u_x(1, y, t) = K_2(y, t), \tag{2.8b}$$

$$u_y(x, 0, t) = K_3(x, t), \tag{2.9a}$$

$$u_y(x, 1, t) = K_4(x, t), \tag{2.9b}$$

with

$$\begin{cases} h_y(x, y) \geqq \gamma > 0, \\ K_{1y}(y, t) < 0, \\ K_{2y}(y, t) > 0, \\ K_3(x, t) \geqq 0, \\ K_4(x, t) \geqq 0. \end{cases} \tag{2.10}$$

Even in this case the regularity follows directly from an estimate on the sign of u_{ny}. Again, this estimate is an immediate consequence of the maximum principle.

3. Uniqueness. In this section we prove an uniqueness result. Let us remark that the right-hand side of (0.1), $\mathbb{H}(*)$ is not Lipschitz continuous and moreover it has the "wrong" sign. Thus the usual techniques cannot be applied (see [9]) and uniqueness turns out not to be a trivial problem.

To make the calculations simpler and to give a more geometrical idea of the techniques applied we will prove uniqueness in a one-dimensional case. (Regularity of the free boundary for the one-dimensional problem follows from [5].)

We assign Dirichlet conditions on the lateral boundary. We stress the fact that this assumption is not essential for our approach.

For the sake of simplicity we introduce the new unknown $u-1$, that we call u again. Let us consider:

$$u_t - u_{xx} = \mathbb{H}(u) \quad \text{in } Q_T = (0, 1) \times (0, T), \tag{3.1}$$

$$u(x, 0) = h(x) \quad \text{in } [0, 1], \tag{3.2}$$

$$u(0, t) = \phi(t) \quad \text{in } [0, T], \tag{3.3}$$

$$u(1, t) = \psi(t) \quad \text{in } [0, T]. \tag{3.4}$$

$$\begin{cases} h(x), \phi(t), \psi(t), \text{ are continuous and } h(x) > 0 \text{ if } x > b, \; h(x) < 0 \\ \text{if } x < b, \; \psi(t) > 0, \; \phi(t) < 0. \text{ Let } u_1(x, t) \text{ and } u_2(x, t) \text{ be} \\ \text{two continuous "weak" solutions in } Q_T \text{ of system (3.1)-(3.4),} \\ \text{(that is, } u_i \in C^0(\bar{Q}_T), \; u_{ixx}, u_{it} \in L_2(Q'_T), \; Q' \subset [0, 1], \; u_{ix} \in C^0(Q_T), \\ i = 1, 2, \text{ and satisfying in a classical sense (3.2)-(3.4) and (3.1) almost} \\ \hfill \text{everywhere).} \\ \text{Let } s_1(t) \text{ and } s_2(t) \text{ represent their level sets } \{u_1 = 0\}, \{u_2 = 0\}, \text{ respectively.} \\ \text{Moreover, let } s_1(t) \text{ and } s_2(t) \text{ be continuous on } [0, T]. \end{cases} \tag{3.5}$$

THEOREM 3.1. *Let (u_1, s_1) and (u_2, s_2) be two solutions of system (3.1)-(3.4) in Q_T, satisfying the assumptions (3.5), and such that either $|u_{1x}| \geqq \gamma > 0$ in a neighbourhood U of s_1 or $|u_{2x}| \geqq \gamma > 0$ in a neighbourhood U of s_2. Then $u_1 \equiv u_2$ and $s_1 \equiv s_2$ in Q_T.*

Proof. Let us put $v := u_2 - u_1$. The function v solves the following problem:

$$\begin{aligned} &v_t - v_{xx} = \mathbb{H}(u_2) - \mathbb{H}(u_1) \quad \text{in } Q_T, \\ &v(x, 0) = v(0, t) = v(1, t) = 0. \end{aligned} \tag{3.6}$$

If $(u_1, s_1) \neq (u_2, s_2)$, we can assume, without loss of generality, that a bifurcation between the two curves $x = s_1(t)$, $x = s_2(t)$ occurs at $t = 0$.

Now $v(x, t)$ can be expressed by means of the Green function $G(x; t; \xi, \tau)$, and we have

$$v(x, t) = \int_0^t d\tau \int_0^1 G(x; t; \xi; \tau) f(\xi, \tau)\, d\xi, \tag{3.7}$$

where $f(x, t) = \mathbb{H}(u_2) - \mathbb{H}(u_1)$, and then $|f(x, t)| = \chi_{Q_{1,2}}$ ($\chi_{Q_{1,2}}$ is the characteristic function of $Q_{1,2} = \{(x, t) \in Q_T\colon s_i(t) \leqq x \leqq s_j(t);\ i, j = 1, 2;\ s_j(t) > s_i(t);\ 0 \leqq t \leqq T\}$). Because of the assumptions we have made, we can also assume that $|u_{1x}| \geqq \gamma > 0$ in $Q_{1,2}$.

Then we have

$$\begin{aligned} |u_1(s_2(t), t) - u_2(s_2(t), t)| &= |u_1(s_2(t), t)| = |u_1(s_2(t), t) - u_1(s_1(t), t)| \\ &\geqq \gamma |s_2(t) - s_1(t)|. \end{aligned}$$

This last inequality is a consequence of the Lagrange mean value theorem and of the assumptions we have made. Hence

$$\begin{aligned} |v(x, t)| &\leqq C \left| \int_0^t \int_0^1 (t-\tau)^{-1/2} \exp\left(\frac{-C(|x-\xi|^2)}{(t-\tau)} \right) \chi_{Q_{1,2}}\, d\xi\, d\tau \right| \\ &\leqq C_1 \int_0^t \frac{|s_1(\tau) - s_2(\tau)|}{\sqrt{t-\tau}}\, d\tau \\ &\leqq C_1 \int_0^t \frac{|u_1(s_2(\tau), \tau) - u_2(s_2(\tau), \tau)|}{\gamma} \frac{1}{\sqrt{t-\tau}}\, d\tau \\ &= C_1 \int_0^t \frac{\|u_1 - u_2\|_{C^0(Q_T)}}{\gamma \sqrt{t-\tau}}\, d\tau. \end{aligned}$$

Then

$$\|u_1 - u_2\|_{C^0(Q_T)} \leqq C_1 \frac{2\sqrt{T}}{\gamma} \|u_1 - u_2\|_{C^0(Q_T)},$$

which for a sufficiently small T gives an absurdum, implying $(u_1, s_1) \equiv (u_2, s_2)$.

Using the same techniques, an equivalent uniqueness theorem can be obtained in the multidimensional case. Let u_1 and u_2 be two "weak" solutions of (1.1)-(1.3d), continuous on $\bar{Q}_T$ and such that their level sets $\{u_1 = 1\}$ and $\{u_2 = 1\}$ have a parametric representation (according to Remark 1.3) $x = f_i(r, t)$, $y = g_i(r, t)$, $i = 1, 2$.

(Let $\hat{S}$ be one between the graphs represented by the parametric surfaces ($x = f_i$; $y = g_i$, $i = 1, 2$), for example, $x = f_1$, $y = g_1$.)

For such solutions the following theorem holds.

THEOREM 3.2. *Let* (u_1, f_1, g_1) *and* (u_2, f_2, g_2) *be two solutions of system* (1.1)-(1.3d) *in* Q_T, *with* u_1 *and* u_2 *continuous on* $\overline{Q_T}$, ∇u_1, $\nabla u_2 \in C^0(\bar{Q}'_T)$, (*with* Q' *as in Theorem* 1.1), f_i *and* g_i *Lipschitz continuous for* $i = 1, 2$; *let, moreover,* $\hat{S}$ *satisfy this condition*: *for all* $P \in \hat{S}$ *at least one of the two assertions holds*:

(a) *There exists an open set* V *such that* $P \in V$; $|u_{1x}| \geqq \gamma > 0$ *in* V *and* $\hat{S}$ *is represented in* V *by a function of the form* $x = F_1(y, t)$;

(b) *There exists an open set* V *such that* $P \in V$; $|u_{1y}| \geqq \gamma > 0$ *in* V *and* $\hat{S}$ *is represented in* V *by a function of the form* $y = G_1(x, t)$; *then* $u_1 \equiv u_2$ *in* Q_T.

(An analogous statement holds if we assume $\hat{S}$ to be the graph represented by $x = f_2$, $y = g_2$.)

We want to stress that this uniqueness theorem implies that the whole sequence of approximating solutions (u_n, f_n) found in § 1 converges, so that the technique displayed in Theorem 1.1 is constructive.

4. Continuous dependence on the data. To fix the basic ideas of the proof, we prove a local continuous dependence theorem for a Dirichlet problem. This means that any solution $u(x, y, t)$, as in Theorem 1.1, depends continuously from the initial and boundary data in a domain Q_{T_1} with T_1 suitably small (naturally T_1 is independent of the particular solution $u(x, y, t)$ chosen). Moreover, we can observe that, in general, if the boundary data satisfy conditions which guarantee the existence of solutions of problem (1.1)-(1.3d) having Lipschitz continuous interfaces in Q_T, represented as functions of the (x, t) variables and such that in their neighbourhood $u_y \geqq \delta > 0$, then our local existence theorem become global (that is, it holds in all Q_T). Other initial boundary value problems can be treated similarly.

THEOREM 4.1. *Let* ${}_1u$ *and* ${}_2u$ *be two functions satisfying* (1.1)-(1.3d) *in the sense specified in Theorem* 1.1, *in* Q_{T_1}, *for a suitable* $T_1 > 0$, *with the same regularity as in Theorem* 1.1. *Let* $y = {}_1f(x, t)$ *and* $y = {}_2f(x, t)$ *be the functions that represent the level sets* $\{{}_1u = 1\}$ *and* $\{{}_2u = 1\}$, *respectively. We assume that* (1.4)-(1.9) *hold and that* ${}_1f$, ${}_2f$ *have the same regularity as in Theorem* 1.1.

Then for each $\hat{\varepsilon} > 0$, $\varepsilon' > 0$ *a* $\delta > 0$ *exists such that if*

(4.1) $$|{}_1h - {}_2h|_{Q}^{(1+\beta)} \leqq \delta,$$

(4.2) $$|{}_1K_i - {}_2K_i|^{(\beta)}_{([0,1]\times[0,T_1])} \leqq \delta, \qquad i = 1, \dots, 4,$$

(4.3) $$|{}_1K_{iy} - {}_2K_{iy}|^{(\beta)}_{([0,1]\times[0,T_1])} \leqq \delta, \qquad i = 1, 2, \beta > \tfrac{3}{4},$$

(4.4) $$\|{}_1K_{it} - {}_2K_{it}\|_{q,((0,1)\times(0,T_1))} \leqq \delta, \qquad i = 1, \dots, 4, \quad q > 4,$$

then

(4.5) $$\|{}_1u - {}_2u\|_{C^0(\bar{Q}_{T_1})} \leqq \hat{\varepsilon},$$

(4.6) $|{}_1u - {}_2u|^{(1+\gamma)}_{(\bar{Q}'_{T_1})} \leqq \hat{\varepsilon}$, *with* $Q' = (0, 1) \times (\varepsilon', 1 - \varepsilon')$ *and* $\gamma = \min(\beta, 1 - 4/q)$,

(4.6bis) $$\|{}_1u - {}_2u\|^{(2)}_{3,Q'_T} \leqq \hat{\varepsilon},$$

(4.7) $\|{}_1u - {}_2u\|^{(2)}_{q,D} \leqq \hat{\varepsilon}$, *where* D *is a domain strictly contained in* Q_T $\forall q > 2$,

(4.7bis) $$\|{}_1u_t - {}_2u_t\|_{C^0(Q'\times[\tau,T_1])} \leqq \hat{\varepsilon},$$

(4.8) $$\|{}_1f - {}_2f\|_{C^0([0,1]\times[0,T_1])} \leqq \hat{\varepsilon},$$

(4.9) $$|{}_1f_x - {}_2f_x|^{(\gamma)}_{([0,1]\times[0,T_1])} \leqq \hat{\varepsilon}.$$

Proof. Let us consider the system of equations satisfied by $w = {}_1u - {}_2u$. We have that $w(x, y, t)$ can be expressed as the sum of two functions w_1, w_2 such that w_1 satisfies the homogeneous heat equation with the same boundary and initial data of w and w_2 satisfies the equation

$$w_{2t} - \Delta w_2 = \mathbb{H}({}_1u - 1) - \mathbb{H}({}_2u - 1) \tag{4.10}$$

with zero boundary and initial data.

As far as w_1 is concerned, we can easily get (by means of a standard maximum principle) that

$$\|w_1\|_{C^0(Q_{T_1})} \leqq \delta. \tag{4.11}$$

We now find an estimate for w_2, solution of (4.10). Using the Green function we get that

$$|w_2(x, y, t)| \leqq \int_0^t d\tau \int_Q |G(x, y, t, \xi_1, \xi_2, \tau)| \chi_{Q_{1,2}} \, d\xi_1 \, d\xi_2, \tag{4.12}$$

where $\chi_{Q_{1,2}}$ is the characteristic function of $Q_{1,2} = \{(x, y, t) \in Q_T\colon {}_if(x, t) \leqq y \leqq {}_jf(x, t)$; $i, j = 1, 2$; ${}_jf(x, t) > {}_if(x, t)$; $0 \leqq x \leqq 1$; $0 \leqq t \leqq T_1\}$. Since we are proving a local continuous dependence, there is no loss of generality in assuming that for a suitable T_1 and sufficiently small γ, both the graphs of ${}_1f(x, t)$ and ${}_2f(x, t)$ lie in a region where ${}_1u_y$, ${}_2u_y \geqq \gamma > 0$.

We have

$$\begin{aligned} |{}_1u(x, {}_2f(x, t), t) - {}_2u(x, {}_2f(x, t), t)| &= |u_1(x, {}_2f(x, t), t) - 1| \\ &= |u_1(x, {}_2f(x, t), t) - u_1(x, {}_1f(x, t), t)| \\ &\geqq \gamma |{}_1f(x, t) - {}_2f(x, t)|. \end{aligned}$$

This last inequality is a consequence of the Lagrange mean value theorem.

Hence we get

$$\|u_1 - u_2\|_{C^0(Q_{T_1})} \geqq \gamma |{}_1f(x, t) - {}_2f(x, t)|. \tag{4.13}$$

This implies that, using (4.12),

$$\|w_2\|_{C^0(Q_t)} \leqq \frac{C}{\gamma} \int_0^t \int_0^1 (t-\tau)^{-1} \exp\left(\frac{-C(|x-\xi_1|^2)}{(t-\tau)}\right) \|{}_1u - {}_2u\|_{C^0(Q_\tau)} \, d\xi_1 \, d\tau.$$

Hence

$$\|w\|_{C^0(Q_t)} \leqq \|w_1\|_{C^0(Q_{T_1})} + C_1 \int_0^1 (t-\tau)^{-3/4} \|w\|_{C^0(Q_T)} \, d\tau.$$

Now, using (4.11), we easily get (4.5), and then, because of (4.13), we also get (4.8).

At this point it is easy to realize that (4.6) holds true, in fact from (4.8) we get that w satisfies a parabolic equation whose source term is bounded and different from zero only between ${}_1f$ and ${}_2f$, that is, in a region whose measure can be made as small as we want choosing a sufficiently small δ. At this point, using inequality (10.12) p. 355 of [6], and the corollary at the end of Theorem 9.1, p. 341 of [6], (4.6) and (4.7) follow immediately, and hence (4.9).

As far as (4.7bis) is concerned, it can be easily obtained using (4.6bis), Proposition 1.4a and Remark 1.4b, since it is simple to prove that two functions which are uniformly Hölder continuous and are "near" in the L_2-norm are also "near" in the C^0-norm.

Using (4.7bis) and (4.8) and setting $W \equiv {}_1u(x, y, \bar{t}) - {}_2u(x, y, \bar{t})$, we have that $W(x, y)$ satisfies an elliptic equation of the kind $\Delta W = \hat{F}_{\bar{t}}$ in Q, where $\|\hat{F}\|_{q,Q''}$, $q \geqq 2$, can be made as small as we want for all $\bar{t} > 0$ and $Q'' \subset \subset Q$ taking δ sufficiently small.

Then, using inequality (10.11) of Theorem 10.1 of [11, p. 173], we have that for all $\bar{t} \in (0, T_1] \| W \|^{(2)}_{2,Q''} \leqq \varepsilon$.

Naturally in this regard δ will depend on $\bar{t}$ and on $d \equiv \text{dist}(Q'', \partial Q)$.

We have proved a local continuous dependence theorem under the assumptions of Theorem 1.1. Actually, following step by step the sketch of the proof, we get continuous dependence theorems in the other cases we have treated in this paper, and in particular we have a global continuous dependence theorem in all the cases in which we have proved a global existence theorem (at least until one of the interfaces of the two solutions touches one of the boundary planes $\{y=0\}$; $\{y=1\}$).

In particular, using the same techniques, we get a local continuous dependence theorem in the case of Remark 1.3.

We want to stress again that the local existence theorem as well as the uniqueness and continuous dependence theorems that we have proved for the two-dimensional case can be obtained in the same way in the general N-dimensional case. Naturally, additional technical complications and some extra assumptions are required to prove global existence theorems in the N-dimensional case.

REFERENCES

[1] J. NORBURY AND A. M. STUART, *A model for porous medium combustion*, Quart. J. Mech. Appl. Math., 42 (1987), pp. 159–178.

[2] ———, *Parabolic free boundary problems arising in porous medium combustion*, IMA J. Appl. Math., 39 (1987), pp. 241–257.

[3] ———, *Travelling combustion waves in a porous medium. Part* 1: *Existence*, SIAM J. Appl. Math., 48 (1988), pp. 155–169.

[4] ———, *Travelling combustion waves in a porous medium. Part* 2: *Stability*, SIAM J. Appl. Math., 48 (1988), pp. 374–392.

[5] R. GIANNI AND P. MANNUCCI, *Existence theorems for a free boundary problem in combustion theory*, Quart. Appl. Math., to appear.

[6] O. A. LADYZHENSKAYA, V. A. SOLONNIKOV, AND V. V. URAL'CEVA, *Linear and quasilinear equations of parabolic type*, Amer. Math. Soc. Transl., 23 (1968).

[7] M. L. KRASNOV, A. I. KISELEV, AND G. I. MAKARENKO, *Equazioni Integrali*, MIR, Moscow, 1976.

[8] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, London, Toronto, 1980.

[9] C. BAIOCCHI AND G. A. POZZI, *An evolution variational inequality related to a diffusion-absorption problem*, Appl. Math. Optim., 2 (1976), pp. 304–314.

[10] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, Heidelberg, New York, 1977.

[11] O. A. LADYZHENSKAYA AND N. N. URAL'CEVA, *Equations aux dérivées partielles de type elliptique*, Monographies Universitaires de Mathématiques, 31, Dunod, Paris, 1968.

SIAM J. MATH. ANAL.
Vol. 24, No. 3, pp. 634–647, May 1993

006

THIN PLATES AND COMPRESSIVE MEMBRANE SOLUTIONS II: A NONEXISTENCE RESULT*

M. E. BREWSTER†

Abstract. It is proved that there is no sequence of solutions of the radially-symmetric von Kármán thin plate equations with non–self-equilibrating pressure load and clamped boundary conditions that has compressive Föppl membrane asymptotics.

Key words. thin plates, compressive membranes, singular perturbations, nonexistence

AMS subject classifications. 34E20, 73K10

1. Introduction.

The deflection of a thin plate can be described by a singular perturbation problem that reduces to the membrane problem for infinitesimal thickness. In [2] formal asymptotic expansions of the plate deflection based on compressive membrane solutions were derived when the edge of the plate is elastically supported against rotation. Also, it was observed that asymptotics based on compressive Föppl membrane theory failed for the clamped plate. In this paper, we present a rigorous proof to support the conjecture that clamped plate solutions do not have compressive Föppl membrane asymptotics.

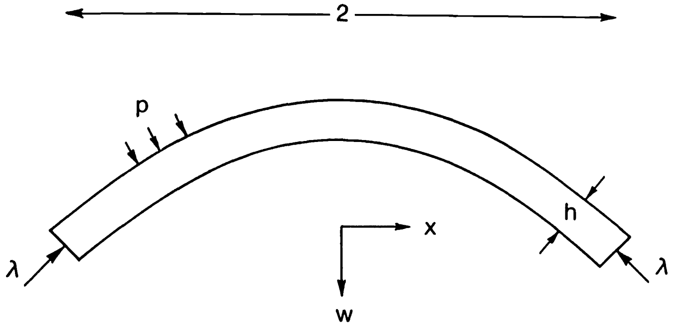


FIG. 1

We consider a thin circular plate subject to a vertical deflection $w(x)$ for $x \in [0, 1]$, where x is radial position (see Fig. 1). The constant h is proportional to the thickness-to-radius ratio. A transverse pressure load $p(x)$ and a radial compressive stress λ at the edge ($x = 1$) are imposed. We consider radially-symmetric solutions where the radial stress $v(x)$ is compressive (positive).

*Received by the editors April 15, 1991; accepted for publication (in revised form) August 14, 1992.

†Program in Applied Mathematics, University of Colorado, Boulder, Colorado 80309-0526. This research was partially supported by Department of Energy grant DE-AS03-76ER72012 and by Army Research Office grant DAAG29-85-K-0092.

The equations of the von Kármán theory (cf. [10]), a nonlinear model that neglects shear stress, are

$$h^2\left(\frac{d^2u}{dx^2}+\frac{3}{x}\frac{du}{dx}\right)+uv=g(x), \tag{1a}$$

$$\frac{d^2v}{dx^2}+\frac{3}{x}\frac{dv}{dx}=u^2, \tag{1b}$$

where

$$u=\frac{1}{x}\frac{dw}{dx}, \qquad g(x)=\frac{1}{hx^2}\int_0^x sp(s)ds.$$

We assume that the pressure load is non–self-equilibrating; that is, $g(1)\neq 0$. The variables may be rescaled according to the magnitude of g; thus, without loss of generality, we assume

$$g(1)=1. \tag{2}$$

The boundary conditions at $x=0$ ensure smoothness at the origin:

$$\frac{dv}{dx}(0)=0, \qquad \frac{du}{dx}(0)=0. \tag{3}$$

At the edge of the plate, we specify the radial stress and also impose the clamped-edge condition:

$$v(1)=\lambda, \tag{4a}$$

$$u(1)=0. \tag{4b}$$

The Föppl membrane equations are obtained by setting $h=0$ in (1a), (1b). From (1a) we get

$$u_0v_0=g(x), \tag{5a}$$

and by eliminating u between (5a) and (1b),

$$\frac{d^2v_0}{dx^2}+\frac{3}{x}\frac{dv_0}{dx}=\frac{g^2(x)}{v_0^2}, \tag{5b}$$

where the zero subscript denotes a membrane solution. We cannot satisfy all of the boundary conditions of the plate problem—we choose

$$\frac{dv_0}{dx}(0)=0, \qquad v_0(1)=\lambda_0. \tag{6}$$

Some background on these problems is given in [1]. The result we will make use of is the existence of compressive ($\lambda>0$) membrane solutions (see [3]).

2. Review of attempts at formal asymptotics. We briefly recall from [2] the attempts to derive formal asymptotic expansions for the clamped plate with compressive radial stress. Let (u_0, v_0) be a compressive solution of the Föppl membrane equations (5). We consider an expansion based on (u_0, v_0):

$$u(x;h) = u_0(x) + x^{-3/2}w(x;h), \tag{7}$$

$$v(x;h) = v_0(x) + v_1(x;h), \tag{8}$$

where we assume $w \ll u_0$, $v_1 \ll v_0$, and $v_0 > 0$. (The w in (7) should not be confused with the vertical deflection.) Then the plate equations (1a), (1b) become

$$h^2\frac{d^2w}{dx^2} + v_0w = -(w + x^{3/2}u_0)v_1 + h^2\left(\frac{3}{4x^2}w - x^{-3/2}\frac{d}{dx}\left(x^3\frac{du_0}{dx}\right)\right), \tag{9}$$

$$\frac{d^2v_1}{dx^2} + \frac{3}{x}\frac{dv_1}{dx} = 2x^{-3/2}wu_0 + x^{-3}w^2. \tag{10}$$

Assuming the right-hand side of (9) is negligible, then the WKBJ approximation is

$$w \sim \frac{c}{v_0^{1/4}}\cos(\tilde{x} + \phi), \qquad c \ll 1, \tag{11}$$

where

$$\tilde{x} = \frac{1}{h}\int^x v_0^{1/2}(s)ds. \tag{12}$$

But the assumption that $c \ll 1$ implies the clamped boundary condition (4b) cannot be satisfied, since

$$u_0(1) = \frac{g(1)}{v_0(1)} = \frac{1}{\lambda_0} \neq 0. \tag{13}$$

We may attempt a WKBJ-like construction for the clamped plate if we drop the assumption that the leading-order term is the Föppl membrane solution. We suppose

$$v(x) \sim v_0(x) + O(h^2)$$

with $v_0 > 0$ and $v_0 = O(1)$ as $h \to 0$, $x > 0$. Then

$$u(x) \sim \frac{g}{v_0} + \frac{\kappa}{x^{3/2}v_0^{1/4}}\cos(\tilde{x} + \phi), \tag{14}$$

where $\tilde{x}$ is defined in (12).

To satisfy the clamped boundary condition (4b), we must have

$$\frac{g(1)}{\lambda} + \frac{\kappa}{\lambda^{1/4}}\cos\theta = 0,$$

where $\theta = \tilde{x} + \phi$ evaluated at $x = 1$. Thus,

$$\kappa = -\frac{g(1)}{\lambda^{3/4}\cos\theta}.$$

Assuming $g(1) \neq 0$, then $\kappa \neq 0$. Observe that the oscillatory term in (14) is $O(1)$.

Upon substitution of (14), (1b) becomes

$$\frac{d^2v}{dx^2} + \frac{3}{x}\frac{dv}{dx} \sim \frac{g^2}{v_0^2} + \frac{\kappa^2}{2x^3 v_0^{1/2}}(\cos(2\tilde{x} + 2\phi) + 1) - \frac{2g\kappa}{x^{3/2}v_0^{5/4}}\cos(\tilde{x} + \phi). \tag{15}$$

The expansion

$$v(x) \sim v_0(x) - h^2 \frac{\kappa^2}{8x^3 v_0^{5/2}} \cos(2\tilde{x} + 2\phi) - h^2 \frac{2g\kappa}{x^{3/2} v_0^{9/4}} \cos(\tilde{x} + \phi)$$

satisfies (15) to leading order provided

$$\frac{d^2v_0}{dx^2} + \frac{3}{x}\frac{dv_0}{dx} = \frac{g^2}{v_0^2} + \frac{\kappa^2}{2x^3 v_0^{1/2}}. \tag{16}$$

It is shown in [2] that (16) has no solutions that are bounded at the origin, and, furthermore, solutions of (16) satisfy

$$v_0'(x) \to -\infty \quad \text{as } x \to 0.$$

Such a function cannot provide a valid asymptotic expansion for solutions of (1), (3), as the formula

$$x^3 v'(x) = \int_0^x s^3 u^2(s) ds$$

demonstrates that $v'(x) > 0$ for all $x > 0$.

3. Statement of nonexistence result. Standard usage of the term "asymptotic" is well understood and precise when used in a positive sense; an asymptotic formula is stated, which may be verified according to, for example, the definition of [4]. Our result, however, is a negative result and it is not so clear what is meant by the nonexistence of an asymptotic expansion based on a given family of outer (membrane) solutions. It is not possible to list or construct all candidates for an expansion; in addition to the classical boundary layer [6], there exist the possibilities of internal layers [8], global breakdown [2], combinations of the above, or perhaps something entirely new. Because of these difficulties, we formulate a definition of "compressive membrane asymptotics" based on the principles of membrane theory.

The stresses employed in membrane theory—the radial stress, v, and the hoopstress $(xv)'$—will be assumed to be bounded (as $h \to 0$). In von Kármán plate theory, the bending stress, $h(u' + (1+\nu)u)$ is included. For membrane theory, the thickness h is zero; thus, the bending stress must be negligible as $h \to 0$. The balance of forces to produce equilibrium, given in the membrane theory by (5a), will be assumed to hold to leading order. To ensure that we are considering compressive membrane states, the radial stress v is assumed positive and bounded away from zero. These conditions are assumed to hold in the least restrictive sense possible—at a point. For purposes of simplifying the proof we will assume that, while the point at which these hypotheses hold is not necessarily fixed, the set of such points must be bounded away from the center of the plate ($x = 0$). At the end of this section, we discuss the relaxation of the last hypothesis.

The conditions described above are incorporated in the following.

DEFINITION. We say that a sequence $\{u_n, v_n, h_n\}$ of solutions of the plate equations (1a), (1b) has compressive membrane asymptotics if

(i) As $n \to \infty$,

$$h_n \to 0; \tag{17}$$

(ii) For some m, $C > 0$, x_0 and points $\{x_n\}$,

$$0 < x_0 < x_n \leq 1, \tag{18a}$$

$$v_n(x_n) \geq m > 0, \tag{18b}$$

$$|v_n(x_n)| + |v_n'(x_n)| \leq C \quad \text{for } n = 1, 2, \dots, \tag{18c}$$

$$|u_n(x_n)v_n(x_n) - g(x_n)| + |h_n u_n'(x_n)| \to 0 \quad \text{as } n \to \infty. \tag{18d}$$

Here is the nonexistence theorem, to be proved in the following sections.

THEOREM 1. *Let $g(x)$ be twice continuously differentiable on $[0,1]$ and $g(1)$ be nonzero. Then there does not exist a sequence of solutions of the clamped plate problem $(1), (3), (4)$ that has compressive membrane asymptotics.*

For technical reasons, the case $x_n \to 0$ was excluded in (18a). We consider here a formal argument to support the conjecture that Theorem 1 would still hold if (18a) were relaxed. The asymptotic formula for solutions of (1a), (1b) derived in [2] is

$$u(x;h) \sim u_0(x) + \frac{c}{x^{3/2} v_0^{1/4}(x)} \sqrt{\frac{\pi}{2}}\, \tilde{x}^{1/2} J_1(\tilde{x}), \tag{19}$$

$$v(x,h) \sim v_0(x), \tag{20}$$

where $\tilde{x}$ is defined in (12). This formula is valid for all $x \in [0,1]$, provided $c \ll h^{1/2}$. Then for $\tilde{x} = O(1)$,

$$|uv - g| + |hu'| \sim O\left(\frac{c}{h^{3/2}}\right)\left(\frac{1}{\tilde{x}} J_1(\tilde{x})\right)' + O\left(\frac{c}{h^{1/2}}\right)\frac{1}{\tilde{x}} J_1(\tilde{x}). \tag{21}$$

Supposing (18d) to hold with $(h_n)^{-1}x_n = O(1)$, then we must have $c \ll h^{1/2}$, at most. Thus (19) gives

$$u(x) - u_0(x) \ll h^{1/2} \quad \text{for } x = O(1), \quad x > 0, \tag{22}$$

which is a *stronger* result than we get in the proof of Theorem 1 as stated. In particular, if $u_0(1) \neq 0$, then the clamped edge condition cannot be satisfied.

In the following two sections, we develop some transformations and lemmas that are needed in the proof of Theorem 1, which is carried out in §6. In §4, we transform the von Kármán plate equations (1a), (1b) into a first-order "fast-slow" system. That is, the variables depending essentially on the "slow" scale (x) are separated from the variables depending on the "fast" scale (x/h). These transformations are motivated by the formal asymptotic expansion for the case of elastically-supported boundary conditions presented in [2]. Lemmas 1 and 2, which are stated and proved in §5, provide a justification of the formal asymptotics (11) when the hypotheses (17), (18) are satisfied. When the clamped edge boundary condition is also assumed to hold, as in Theorem 1, then (13) provides a contradiction.

4. Transformations. We consider the deviation of the plate solutions from a given compressive membrane solution (u_0, v_0) satisfying the Föppl membrane equations (5) on $(\xi_0, 1]$ with $\xi_0 \geq 0$ and v_0 positive on $(\xi_0, 1]$. Following the formal asymptotics above, and motivated by transformations in [7], we define

$$u_1(t) = v_0^{1/4} x^{3/2}(u - u_0), \qquad v_1(t) = v_0^{1/4} x^{3/2}(v - v_0), \tag{23}$$

where

$$t = \int_{\xi_1}^{x} v_0^{1/2}(s)ds \tag{24}$$

with $\xi_0 < \xi_1 \leq 1$. The interval $(\xi_0, 1]$ maps onto $(T_0, T_1]$ under the change of variables (24), where $T_0 \geq -\infty$ and $0 \leq T_1 < +\infty$. Upon substitution into (1a), (1b), we have

$$h^2 \frac{d^2 u_1}{dt^2} + u_1 + \beta v_1 = -\gamma u_1 v_1 + h^2 q - h^2 \alpha u_1, \tag{25a}$$

$$\frac{d^2 v_1}{dt^2} - 2\beta u_1 + \alpha v_1 = \gamma u_1^2, \tag{25b}$$

for $t \in (T_0, T_1]$, where $\alpha(t), q(t)$, and $\gamma(t) \in C(T_0, T_1]$, and

$$\beta(t) = \frac{g(x)}{v_0^2(x)} \in C^1(T_0, T_1]. \tag{26}$$

Formulas for α, q, and γ are given in Appendix A.

We now convert these two second-order equations into a first-order "fast-slow" system by the substitution of

$$\mathbf{z}_I = \begin{pmatrix} u_1 + \beta v_1 \\ h\frac{d}{dt}u_1 \end{pmatrix}, \qquad \mathbf{z}_{II} = \begin{pmatrix} v_1 \\ \frac{d}{dt}v_1 + 2h^2\beta\frac{d}{dt}u_1 \end{pmatrix} \tag{27}$$

into (26a), (26b) to give

$$h\frac{d}{dt}\mathbf{z}_I = B\left(\mathbf{z}_I + Q(t, \mathbf{z}_I, \mathbf{z}_{II})\right) + hF_I(t, \mathbf{z}_I, \mathbf{z}_{II}; h), \tag{28a}$$

$$\frac{d}{dt}\mathbf{z}_{II} = A\mathbf{z}_{II} + F_{II}(t, \mathbf{z}_I, \mathbf{z}_{II}; h), \tag{28b}$$

where $A(t)$ is continuous on $(T_0, T_1]$,

$$B = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad \text{and} \quad Q(t, \mathbf{x}, \mathbf{y}) = \begin{pmatrix} \gamma(x_1 - \beta y_1)y_1 \\ h\left(\frac{d\beta}{dt}y_1 + \beta y_2\right) \end{pmatrix}. \tag{29}$$

The $O(h^0)$ linear terms in (28a), (28b) are $B\mathbf{z}_I$ and $A\mathbf{z}_{II}$, respectively; thus the "fast" system (28a) and the "slow" system (28b) decouple at this level. Quadratic terms are contained in Q and F_{II}. Also, terms that are $o(h^0)$ appear in F_I, F_{II}, and Q. The formulas for these terms are given in Appendix A.

5. Properties of the fast-slow system. In this section, we derive some results concerning the asymptotic behavior of solutions of a class of fast-slow systems that contains (28) as a special case. Consider the system

$$h\frac{d}{dt}\mathbf{z}_I = B\left(\mathbf{z}_I + \nabla_I\phi(t,\mathbf{z}_I,\mathbf{z}_{II};h)\right) + hF_I(t,\mathbf{z}_I,\mathbf{z}_{II};h), \tag{30a}$$

$$\frac{d}{dt}\mathbf{z}_{II} = A(t)\mathbf{z}_{II} + F_{II}(t,\mathbf{z}_I,\mathbf{z}_{II};h), \tag{30b}$$

where ∇_I represents the gradient with respect to the "fast" components, $\mathbf{z}_I$. The vectors $\mathbf{z}_I, \mathbf{z}_{II}$ are of arbitrary finite dimension, and the matrix B is assumed to be antisymmetric, which implies the eigenvalues are pure imaginary. To ensure the existence of solutions of (30), the right-hand side of (30) is assumed to depend continuously on $t \in [\tau_0, T_1]$ and $\mathbf{z}_I, \mathbf{z}_{II}$ whenever $|\mathbf{z}_I| + |\mathbf{z}_{II}| < C_0$ for some $C_0 > 0$. The notation $|\cdot|$ represents the Euclidean vector norm. Further, suppose that for $J = I, II$,

$$F_J(t,\mathbf{z}_I,\mathbf{z}_{II},h) = O\left((|\mathbf{z}_I| + |\mathbf{z}_{II}|)^2\right) + o(h^0) \quad \text{as } h, |\mathbf{z}_I|, |\mathbf{z}_{II}| \to 0; \tag{31}$$

that is, there exist positive constants C_1, C_2 such that

$$|F_J(t,\mathbf{z}_I,\mathbf{z}_{II},h)| \le C_1\left((|\mathbf{z}_I| + |\mathbf{z}_{II}|)^2\right) + \psi(h)$$

whenever $t \in [\tau_0, T_1]$ and $h + |\mathbf{z}_I| + |\mathbf{z}_{II}| < C_2$, where

$$\psi(h) \downarrow 0 \quad \text{as } h \to 0.$$

Also, suppose that

$$\begin{gathered} \phi = O\left((|\mathbf{z}_I| + |\mathbf{z}_{II}|)^3\right) + o(h^0), \\ \frac{\partial\phi}{\partial t} = O\left((|\mathbf{z}_I| + |\mathbf{z}_{II}|)^3\right) + o(h^0), \\ \nabla_J\phi = O\left((|\mathbf{z}_I| + |\mathbf{z}_{II}|)^2\right) + o(h^0) \end{gathered} \tag{32}$$

as $h, |\mathbf{z}_I|, |\mathbf{z}_{II}| \to 0, \quad J = I, II$.

We begin with a formal asymptotic study of (30) assuming $h \to 0$ and that the solution is also small. Thus we neglect nonlinear terms and terms that are $o(h^0)$ to obtain

$$h\frac{d}{dt}\mathbf{z}_I = B\mathbf{z}_I, \tag{33a}$$

$$\frac{d}{dt}\mathbf{z}_{II} = A(t)\mathbf{z}_{II}. \tag{33b}$$

Given functions $\mathbf{z}_I, \mathbf{z}_{II}$ on $[\tau_0, T_1]$, we define a scalar function

$$\mathcal{A}(t) = \left(|\mathbf{z}_I(t)|^2 + |X^{-1}(t)\mathbf{z}_{II}(t)|^2\right)^{1/2}, \tag{34}$$

where X is a fundamental matrix solution of (33b) with $X(T_1) = I$. Because A is continuous on $[\tau_0, T_1]$, then X is invertible on this interval and $\mathcal{A}$ is well defined. If $(\mathbf{z}_I, \mathbf{z}_{II})$ is a solution of (30), then

$$\frac{1}{2}\frac{d}{dt}\mathcal{A}^2 = \frac{1}{h}\mathbf{z}_I^T B\mathbf{z}_I + \mathbf{z}_{II}^T X^{-1T}\left((X^{-1})'\mathbf{z}_{II} + X^{-1}A\mathbf{z}_{II}\right). \tag{35}$$

Because B is antisymmetric, $\mathbf{z}_I^T B \mathbf{z}_I = 0$. Also, from

$$X' = AX,$$

we obtain

$$(X^{-1})' = -X^{-1}A.$$

Hence,

$$\frac{d}{dt}\mathcal{A}^2 = 0;$$

that is, $\mathcal{A}$ is a constant.

Next, we consider the full system (30). Let $(\mathbf{z}_I, \mathbf{z}_{II})$ be a solution of (30) and define $\mathcal{A}$ by (34) as before. Then

$$\frac{1}{2}\frac{d}{dt}\mathcal{A}^2 = \mathbf{z}_I^T \frac{d}{dt}\mathbf{z}_{II} + (X^{-1}\mathbf{z}_{II})^T X^{-1} F_{II}. \tag{36}$$

Now

$$\mathbf{z}_I^T \frac{d}{dt}\mathbf{z}_I = (\mathbf{z}_I + \nabla_I \phi)^T \frac{d}{dt}\mathbf{z}_I - \nabla_I \phi^T \frac{d}{dt}\mathbf{z}_I,$$

and substitution from (30a) gives

$$\begin{aligned}\mathbf{z}_I^T \frac{d}{dt}\mathbf{z}_I &= (\mathbf{z}_I + \nabla_I \phi)^T \left(\frac{1}{h}\, B(\mathbf{z}_I + \nabla_I \phi) + F_I\right) - \nabla_I \phi^T \frac{d}{dt}\mathbf{z}_I,\\ &= F_I^T(\mathbf{z}_I + \nabla_I \phi) - \nabla_I \phi^T \frac{d}{dt}\mathbf{z}_I,\end{aligned}$$

where we have used the assumption that B is antisymmetric to eliminate the $O(h^{-1})$ term. Thus,

$$\frac{1}{2}\frac{d}{dt}\mathcal{A}^2 = F_I^T(\mathbf{z}_I + \nabla_I \phi) - \nabla_I \phi^T \frac{d}{dt}\mathbf{z}_I + (X^{-1}\mathbf{z}_{II})^T X^{-1} F_{II}. \tag{37}$$

Integrating from τ to t gives

$$\mathcal{A}^2(t) - \mathcal{A}^2(\tau) = -2(\phi(t, \mathbf{z}_I(t), \mathbf{z}_{II}(t); h) - \phi(\tau, \mathbf{z}_I(\tau), \mathbf{z}_{II}(\tau); h)) + 2\int_\tau^t f(s)ds, \tag{38}$$

where

$$\begin{aligned} f(t) =& F_I^T\,(t, \mathbf{z}_I(t), \mathbf{z}_I(t); h)\,(\mathbf{z}_I(t) + \nabla_I \phi(t, \mathbf{z}_I(t), \mathbf{z}_{II}(t); h))\\ &+ \frac{\partial \phi}{\partial t}(t, \mathbf{z}_I(t), \mathbf{z}_{II}(t); h)\\ &+ \nabla_{II}\phi(t, \mathbf{z}_I(t), \mathbf{z}_{II}(t); h)\,(A(t)\mathbf{z}_{II}(t) + F_{II}(t, \mathbf{z}_I(t), \mathbf{z}_{II}(t); h))\\ &+ (X^{-1}(t)\mathbf{z}_{II}(t))^T\, X^{-1}(t) F_{II}(t, \mathbf{z}_I(t), \mathbf{z}_{II}(t); h).\end{aligned} \tag{39}$$

We now derive from the integral equation (38) an integral inequality for $\mathcal{A}(t)$. First, note that since $X^{-1}(t)$ is bounded on $[\tau_0, T_1]$, then the definition of $\mathcal{A}$, (34), implies

$$\mathbf{z}_I(t) = O(\mathcal{A}(t)) \quad \text{and} \quad \mathbf{z}_{II}(t) = O(\mathcal{A}(t)) \tag{40}$$

for $t \in [\tau_0, T_1]$. From (40), hypotheses (31), (32) and the integral equation (38), there exist positive constants C_1, C_2, independent of h, such that

$$\mathcal{A}^2(t) - \mathcal{A}^2(\tau) \leq C_1 \left(\mathcal{A}^3(t) + \mathcal{A}^3(\tau)\right) + C_1 \int_\tau^t \mathcal{A}^3(s)ds + \psi(h) \tag{41}$$

whenever

$$h + \mathcal{A}(s) \leq C_2 \tag{42}$$

for all $s \in [\tau, t]$ with $\tau_0 \leq \tau \leq t \leq T_1$, where

$$\psi(h) \downarrow 0 \quad \text{as } h \to 0.$$

We now present Lemma 1, which will be used in the proof of our second lemma on the validity of the formal asymptotic expansion (7), (8). The proof of Lemma 1 is given in Appendix B.

LEMMA 1. *Let $z(s) \geq 0$ be a continuous function defined on $s \in [0, 1]$, satisfying*

$$z^2(s) - z^2(0) \leq z^3(s) + z^3(0) + \int_0^s z^3(r)dr + \epsilon, \tag{43}$$

where $\epsilon \geq 0$. Then there exist $\delta_0 > 0$ and $k_0 > 0$ such that $z(s) \leq 4\delta$ for $s \in [0, 1]$ whenever $0 \leq z(0) \leq \delta \leq \delta_0$ and $\epsilon \leq k_0\delta^2$.

We will now apply Lemma 1 to obtain estimates for $\mathcal{A}$. Consider $h \in (0, C_2/2)$ and suppose $\mathcal{A}(\tau) \leq C_2/8$ for some $\tau \in [\tau_0, T_1]$. Because $\mathcal{A}$ is continuous, there is a largest value T_2 in $(\tau, T_1]$ such that (42) is satisfied on $[\tau, T_2]$. Thus, (41) holds on $[\tau, T_2]$. Let

$$\begin{aligned} m &= \max\{1, T_1 - \tau_0\}, & s &= \frac{t - \tau}{T_2 - \tau} \\ z(s) &= mC_1\mathcal{A}(t), & \epsilon &= m^2C_1^2\psi(h). \end{aligned} \tag{44}$$

Then, from (41),

$$\begin{aligned} z^2(s) - z^2(0) &= m^2C_1^2(\mathcal{A}^2(t) - \mathcal{A}^2(\tau)) \\ &\leq m^2C_1^3(\mathcal{A}^3(t) + \mathcal{A}^3(\tau)) + m^2C_1^3 \int_\tau^t \mathcal{A}^3(p)dp + m^2C_1^2\psi(h) \\ &\leq \frac{1}{m}(z^3(s) + z^3(0)) + \frac{T_2 - \tau}{m} \int_0^s z^3(r)dr + \epsilon \\ &\leq z^3(s) + z^3(0) + \int_0^s z^3(r)dr + \epsilon \end{aligned}$$

for $s \in [0, 1]$. From Lemma 1, if for some $\delta > 0$, we have

$$z(0) = mC_1\mathcal{A}(\tau) \leq \delta \leq \delta_0, \tag{45}$$

$$\epsilon = m^2C_1^2\psi(h) \leq k_0\delta^2, \tag{46}$$

then

$$z(s) = mC_1\mathcal{A}(t) \leq 4\delta \quad \text{for } t \in [\tau, T_2]. \tag{47}$$

If $\mathcal{A}(\tau) \leq \delta_0/mC_1$, then we may choose δ such that (45) is satisfied and

$$\delta \leq \frac{C_2}{8}mC_1.$$

Thus, (47) implies $\mathcal{A}(T_2) \leq C_2/2$. If $T_2 < T_1$, then there exists $T_2^* > T_2$ such that (42) holds on $[\tau, T_2^*]$, contradicting the assumption that T_2 is the largest such value. Hence $T_2 = T_1$, and (47) holds for $t \in [\tau, T_1]$. We now have the following.

LEMMA 2. *Suppose that* (31), (32) *are satisfied by the fast-slow system* (30) *defined on* $[\tau_0, T_1]$. *There exist a constant* $\delta_1 > 0$ *and a positive function* $h_1(\delta)$ *such that if*

$$0 \le \delta \le \delta_1 \quad \textit{and} \quad 0 < h < h_1(\delta), \tag{48}$$

$(\mathbf{z}_I, \mathbf{z}_{II})$ *is a solution, in a neighborhood of* $\tau \in [\tau_0, T_1]$, *of* (30) *for this value of* h, $\mathcal{A}(t)$ *is defined as in* (34) *and*

$$\mathcal{A}(\tau) \le \delta, \tag{49}$$

then $(\mathbf{z}_I, \mathbf{z}_{II})$ *can be extended to the interval* $[\tau, T_1]$, *and on this interval*

$$\mathcal{A}(t) \le 4\delta. \tag{50}$$

Proof. Let

$$0 < \delta_1 < \min\left(\frac{C_2}{8}, \frac{\delta_0}{mc_1}\right) \quad \text{and} \quad 0 < h_1(\delta) < \min\left(\frac{C_2}{2}, \tilde{h}\right), \tag{51}$$

where $\psi(\tilde{h}) \le k_0\delta^2$, $\tilde{h} > 0$. Such an $\tilde{h}$ exists, since $\psi(h) \to 0$ as $h \to 0$. Then, as in the discussion above, (48), (49) imply that (50) holds on the interval of existence (to the right) of the solution. Therefore, the solution values are finite at the right-hand endpoint of the interval of existence, which must be the entire interval $[\tau, T_1]$ (cf. [4], p. 15). □

The proof of this result was motivated to some extent by [9]. There are some important differences in the result of [9] and our Lemma 2. We do not require the initial condition $\mathcal{A}(\tau)$ to be $O(h)$, and the bounds given in Lemma 2 apply only for $O(1)$ intervals. The form of the system (30), particularly with respect to nonlinear terms, is more restrictive than the systems considered in [9].

6. Proof of Theorem 1.

We now present the proof of the nonexistence result (Theorem 1), which was stated in §2. The proof is by contradiction; thus, suppose there exists a sequence $\{u_n, v_n, h_n\}$, $n = 1, 2, \ldots$ satisfying the clamped plate problem (1), (3), (4) for some sequence of edge loads, $\{\lambda_n = v_n(1)\}$, and further suppose that this sequence has compressive membrane asymptotics; that is, (17), (18) are satisfied. We may extract a subsequence such that for some constants v_∞, v'_∞ and for some $\xi_1 \in [x_0, 1]$,

$$x_{n_k} \to \xi_1, \quad v_{n_k}(x_{n_k}) \to v_\infty \quad \text{and} \quad v'_{n_k}(x_{n_k}) \to v'_\infty \tag{52}$$

as $k \to \infty$. To simplify notation, we drop the repeated subscript. Hypothesis (18) implies $v_\infty \ge m > 0$. Also, one integration of (1b) gives

$$\frac{dv}{dx}(x) = x^{-3} \int_0^x s^3 u^2(s)\,ds \ge 0;$$

hence $v'_\infty \ge 0$.

Let (u_0, v_0) be the solution of the Föppl membrane equations (5a), (5b) satisfying

$$v_0(\xi_1) = v_\infty, \qquad v'_0(\xi_1) = v'_\infty. \tag{53}$$

It is shown in Appendix C that v_0 exists and is positive on an interval $(\xi_0, 1]$ with

$$x_0 \leq \xi_0 < \xi_1.$$

There exists an integer $N \geq 1$ such that

$$\xi_0 < x_n \quad \text{for } n \geq N.$$

Because $v_0 \in C^1(\xi_0, 1]$ and from (52), we have

$$\text{(54a)} \qquad v_n(x_n) - v_0(x_n) \to 0, \quad v_n'(x_n) - v_0'(x_n) \to 0 \quad \text{as } n \to \infty.$$

Further, $u_0 \in C^1(\xi_0, 1]$ from the assumption that g is continuously differentiable. Thus, $\{u_0'(x_n)\}$ is a bounded sequence, and hypotheses (17), (18b), (18d) imply

$$\text{(54b)} \qquad u_n(x_n) - u_0(x_n) \to 0 \quad \text{and} \quad h_n(u_n'(x_n) - u_0'(x_n)) \to 0 \quad \text{as } n \to \infty.$$

We now apply the transformations (23)–(27) to each solution $(u_n(x), v_n(x))$ with corresponding thickness h_n to obtain a sequence $\mathbf{z}_{In}(t)$, $\mathbf{z}_{IIn}(t)$ of solutions of the fast-slow system (28). According to the definition (24) of t, $x = \xi_0$, ξ_1 and 1 correspond to $t = T_0$, 0, and T_1, respectively. Let the sequence $\{t_n\} \in (T_0, T_1], n \geq N$ correspond to $\{x_n\}$ as in definition (24). Then by (52), $t_n \to 0$ as $n \to \infty$. From (23), (27), (54), we have

$$\mathbf{z}_{In}(t_n), \mathbf{z}_{IIn}(t_n) \to 0 \quad \text{as } n \to \infty.$$

Let $\tau_0 = \inf_{n \geq N} t_n$. The fundamental matrix solution $X(t)$ of (33b) is independent of n and its inverse is bounded uniformly for $t \in [\tau_0, T_1]$. Hence, from (34),

$$\text{(55)} \qquad \mathcal{A}_n(t_n) \to 0 \quad \text{as } n \to \infty.$$

We now apply Lemma 2. It is easy to show that the hypotheses (31), (32) are satisfied on $[\tau_0, T_1]$ for the system (28), where the elements of (28) are defined in Appendix A. By (17), (55), and Lemma 2,

$$\text{(56)} \qquad \mathcal{A}_n(T_1) \to 0 \quad \text{as } n \to \infty.$$

We defined $X(T_1)$ as the identity matrix, so from (35) and (56) we have

$$|\mathbf{z}_{In}(T_1)|^2 + |\mathbf{z}_{IIn}(T_1)|^2 \to 0 \quad \text{as } n \to \infty.$$

From (27), we have

$$|u_{1n}(T_1) + \beta(T_1)(\lambda_n - \lambda_0)| + |\lambda_n - \lambda_0| \to 0 \quad \text{as } n \to \infty.$$

Hence, $\lambda_n \to \lambda_0$ and $u_{1n}(1) \to 0$ as $n \to \infty$. But, by (23),

$$u_{1n}(T_1) = \lambda_0^{1/4}(u_n(1) - u_0(1)).$$

The clamped edge assumption gives $u_n(1) = 0$, so

$$u_{1n}(T_1) = -\frac{g(1)}{\lambda_0^{3/4}} \neq 0$$

by the assumption of non–self-equilibrating load, i.e., $g(1) \neq 0$. We have arrived at a contradiction, and thus the assumption that a sequence of clamped plate solutions exist that asymptotically approach the family of compressive membrane solutions must be false. □

7. Conclusion. The definition of compressive membrane asymptotics appears to be rather weak, but, as shown in Lemma 2, guarantees strong (uniform) convergence of a subsequence to some membrane solution. This convergence provides a contradiction to the clamped boundary condition, which is not satisfied by the membrane solution.

Referring to our earlier discussion of the formal asymptotics in §2, one could conjecture that clamped plate solutions also do not converge *weakly* to compressive membrane solutions. This conjecture has not been addressed in Theorem 1, as the definition of asymptotic approach is too restrictive to allow weak convergence. The extension of the proof of Theorem 1 to the case of weak convergence is under investigation.

The clamped circular plate problem provides a striking example of a singular perturbation problem where some solutions of the "outer" problem have no counterpart in the original problem. We have presented here a rigorous justification of an aspect of this phenomenon.

Appendix A. Some formulas from the transformations discussed in §3 are given below. Let

$$\mathcal{L} = \frac{d^2}{dx^2} + \frac{3}{x}\frac{d}{dx}.$$

Then

$$\begin{aligned}
\alpha(t) &= \frac{x^{3/2}}{v_0^{3/4}(x)}\,\mathcal{L}\left(x^{-3/2}\,v_0^{-1/4}(x)\right), \qquad q(t) = -\frac{x^{3/2}}{v_0^{3/4}(x)}\,\mathcal{L}\,u_0(x),\\
\beta(t) &= \frac{g(x)}{v_0^2(x)}, \qquad \gamma(t) = x^{-3/2}v_0^{-5/4}(x),\\
A(t) &= \begin{pmatrix} 0 & 1\\ -\alpha - 2\beta^2 & 0\end{pmatrix},\\
F_I(t,\mathbf{x},\mathbf{y};h)_1 &= -2h\beta^2 x_2,\\
F_I(t,\mathbf{x},\mathbf{y};h)_2 &= -h\alpha x_1 + h\alpha\beta y_1 + hq,\\
F_{II}(t,\mathbf{x},\mathbf{y};h)_1 &= -2h\beta x_2,\\
F_{II}(t,\mathbf{x},\mathbf{y};h)_2 &= -2h^2\alpha\beta x_1 + 2h\,\frac{d\beta}{dt}x_2 + 2h^2\beta q\\
&\quad + \gamma(x_1 - 3\beta y_1)(x_1 - \beta y_1) + 2h^2\alpha\beta^2 y_1.
\end{aligned}$$

For the system in form (31), we have

$$\phi(t,\mathbf{x},\mathbf{y}) = \frac{\gamma}{2}(x_1 - 2\beta y_1)x_1 y_1 + h\frac{d\beta}{dt}x_2 y_1 + h\beta x_2 y_2.$$

This gives

$$\nabla_I \phi = Q.$$

Appendix B.

Proof of Lemma 1. Suppose

$$0 < \delta \le \delta_0 \quad \text{and} \quad 0 \le \epsilon \le k_0\delta^2,$$

where $\delta_0 = 1/14$ and $k_0 = 37/14$. We first show that there exists a continuous function $\bar{z}(s)$ that satisfies the integral equation

$$\bar{z}^2(s) - \delta^2 = \bar{z}^3(s) + \delta^3 + \int_0^s \bar{z}^3(r)dr + \epsilon, \tag{57}$$

for $s \in [0,1]$, $\bar{z} \in [\delta, 4\delta]$. At $s = 0$, (57) becomes

$$\bar{z}^3(0) - \bar{z}^2(0) + \delta^2 + \delta^3 + \epsilon = 0. \tag{58}$$

Let $P(\zeta) = \zeta^3 - \zeta^2 + \delta^2 + \delta^3 + \epsilon$. Clearly, $P(\delta) > 0$. Also,

$$P(2\delta) = 5\delta^3 - 3\delta^2 + \epsilon \le \delta^2(5\delta_0 - 3 + k_0) = 0. \tag{59}$$

Thus, P has a zero in $(\delta, 2\delta]$, say $\bar{z}_0$. Let $\bar{z}(s)$ be the solution of the initial-value problem

$$(2\bar{z} - 3\bar{z}^2)\bar{z}' = \bar{z}^3, \qquad \bar{z}(0) = \bar{z}_0. \tag{60}$$

Then $\bar{z}$ is also a solution of (57). Solving (60) gives $\bar{z}$ implicitly as

$$s = \frac{2(\bar{z}(s) - \bar{z}_0)}{\bar{z}_0 \bar{z}(s)} - 3 \log \frac{\bar{z}(s)}{\bar{z}_0}. \tag{61}$$

For $\bar{z} \in [\delta, 4\delta]$ we have

$$\frac{ds}{d\bar{z}} = \frac{2 - 3\bar{z}}{\bar{z}^2} > 0. \tag{62}$$

Hence, the inverse function theorem guarantees the existence of $\bar{z}(s)$ on some interval $[0, s(4\delta)]$. To estimate the size of this interval, (61) gives

$$s(4\delta) \ge \frac{1}{2\delta_0} - 3\log 4 \ge 1. \tag{63}$$

Thus (61) can be inverted to give a single-valued function $\bar{z}(s)$ for $s \in [0,1]$, $\bar{z} \in [\delta, 4\delta]$ satisfying (57).

We now show that $\bar{z}(s)$ is an upper bound for any positive, continuous solution of

$$z^2(s) - z^2(0) \le z^3(s) + z^3(0) + \int_0^s z^3(r)dr + \epsilon \tag{64}$$

for $s \in [0,1]$ with $z(0) \le \delta < \bar{z}(0)$. Suppose this is not true. Then by continuity of z and $\bar{z}$, there is a smallest value, $s^* > 0$, such that $z(s^*) = \bar{z}(s^*)$. Thus,

$$\bar{z}(s) > z(s) > 0 \quad \text{for } 0 \le s < s^*. \tag{65}$$

The integral (in)equalities (57), (64) give

$$\begin{aligned}\bar{z}^2(s^*) &= \bar{z}^3(s^*) + \delta^2 + \delta^3 + \int_0^{s^*} \bar{z}^3(r)dr + \epsilon \\ &> z^3(s^*) + z(0)^2 + z(0)^3 + \int_0^{s^*} z^3(r)dr \ge z^2(s^*),\end{aligned}$$

contradicting that $\bar{z}(s^*) = z(s^*)$. Hence,

$$z(s) < \bar{z}(s) \le 4\delta \quad \text{for } 0 \le s \le 1,$$

as required for the proof of Lemma 1. □

Appendix C. The following existence-uniqueness result is required for the proof of Theorem 1.

LEMMA 3. *Let g be continuous on* $[0,1]$. *Then the solution of the Föppl membrane equation*

$$v'' + \frac{3}{x}v' = \frac{g^2}{v^2}, \qquad x \in [0,1] \tag{66}$$

with initial conditions

$$v(\xi_1) = v_\infty > 0, \qquad v'(\xi_1) = v'_\infty \geq 0 \tag{67}$$

has a unique solution that exists for $x \in (\xi_0, 1]$, *where* $\xi_0 < \xi_1 < 1$.

Proof. By the standard results [4], there exists a unique solution of (66), (67) on an interval $J = (\xi_0, \xi_2)$, with $\xi_1 \in J$. Integrating (66) once gives

$$x^3 v'(x) = \xi_1^3 v'(\xi_1) + \int_{\xi_1}^{x} s^3 \frac{g^2(s)}{v(s)} ds.$$

Thus $v'(x) \geq 0$ for $x \in J$. Further,

$$v(x) \geq v_\infty > 0 \quad \text{for } x \geq \xi_1.$$

The right-hand side of (66) is bounded and satisfies a Lipschitz condition with respect to v for $v \geq v_\infty$. Thus by the Picard–Lindelöf theorem [5], the solution exists and is unique on $(\xi_0, 1]$. □

Acknowledgment. The author wishes to thank H. B. Keller for suggesting this problem.

REFERENCES

[1] M. E. BREWSTER, *Asymptotic Analysis of Thin Plates Under Normal Load and Horizontal Edge Thrust*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 1987.

[2] ———, *Thin plates and compressive membrane solution* I: *global breakdown*, SIAM J. Appl. Math., 51 (1991), pp. 1255–1283.

[3] A. J. CALLEGARI, E. L. RIESS, AND H. B. KELLER, *Membrane buckling: a study in solution multiplicity*, Comm. Pure Appl. Math., 24 (1971), pp. 499–527.

[4] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill Book Company, New York, 1955.

[5] A. ERDELYI, *Asymptotic Expansions*, Dover Publications, New York, 1956.

[6] K. O. FRIEDRICHS AND J. J. STOKER, *The non-linear boundary value problem of the buckled plate*, Amer. J. Math., 63 (1941), pp. 839–888.

[7] N. FRÖMAN AND P. O. FRÖMAN, *JWKB Approximation; Contributions to the Theory*, North-Holland Publishing Company, Amsterdam, 1965.

[8] J. KEVORKIAN AND J. D. COLE, *Perturbation Methods in Applied Mathematics*, Springer-Verlag, New York, 1981.

[9] H.-O. KREISS, *Problems with different time scales for ordinary differential equations*, SIAM J. Numer. Anal., 16 (1979), pp. 980–998.

[10] J. J. STOKER, *Nonlinear Elasticity*, Nelson, London, 1968.

SIAM J. MATH. ANAL.
Vol. 24, No. 3, pp. 648-668, May 1993

007

SOME GENERAL EXISTENCE PRINCIPLES AND RESULTS FOR $(\phi(y'))=qf(t,y,y'), 0<t<1$*

DONAL O'REGAN†

Abstract. Existence principles and results are reported for the second-order differential equation $(\phi(y'))'=qf(t,y,y')$, $0<t<1$ with y satisfying Dirichlet or mixed boundary data. In particular the case $\phi(v)=|v|^{n-2}v$, $n>1$ is included.

Key words. p-Laplacian, boundary value problems, existence principles

AMS subject classification. 34B15

1. Introduction. In this article existence results are presented for the second-order differential equation

$$(1.1) \qquad (|u'|^{n-2}u')'=q(t)f(t,u,u'), \quad 0<t<1 \quad \text{and} \quad n\in\mathbf{R}, \quad n>1$$

with u satisfying either the Dirichlet boundary condition

(i) $u(0)=a, u(1)=b$

or the mixed boundary condition

(ii) $u'(0)=a, u(1)=b.$

Equations of the above form occur in the study of the n-Laplace equation [9], non-Newtonian fluid theory [8], and the turbulent flow of a gas in a porous medium [1], [6]. The existence results obtained in this paper will improve, extend, and compliment the existing theory found in [1], [4], [9]. In fact even new results are obtained for the case $n=2$.

We summarize briefly known results for the Dirichlet boundary value problem when $n=2$. In [5], [7] it was shown that if f satisfies the monotonicity condition

$$uf(t,u,0)>0 \quad \text{for } |u|>M \quad \text{and} \quad t\in[0,1],$$

together with the growth condition (Bernstein-Nagumo)

$$|f(t,u,p)|\leqq \psi(|p|) \text{ for } (t,u)\in[0,1]\times[-M,M],$$

where $\psi\colon[0,\infty)\to(0,\infty)$ is a continuous function and

$$2M<\int_0^\infty \frac{x\,dx}{\psi(x)},$$

then

$$(*) \qquad \begin{aligned} &y''=f(t,y,y'), \quad 0<t<1, \\ &y(0)=y(1)=0 \end{aligned}$$

has a $C^2[0,1]$ solution. The technique (which uses arguments based on the supremum norm) is referred to as the Bernstein-Nagumo theory in the literature.

On the other hand, suppose $f(t,u,p)=f_1(t,u,p)+f_2(t,u,p)$ and the monotonicity condition above is replaced by

$$uf_1(t,u,p)\geqq a|u|^2 \quad \text{for } (t,u,p)\in[0,1]\times\mathbf{R}^2,$$

* Received by the editors November 20, 1991; accepted for publication (in revised form) August 31, 1992.

† Department of Mathematics, University College Galway, Galway, Ireland.

together with

$$|f_2(t, u, p)| \leqq K\{|u|^\alpha + |p|^\beta + 1\} \quad \text{for } 0 \leqq \alpha, \beta < 1.$$

The results in [12], [13] imply that if f_1 satisfies a growth condition of the form

$$|f_1(t, u, p)| \leqq Ap^2 + B \quad \text{with } A, B \text{ constants},$$

then $(*)$ has a solution provided $a > -\pi^2$. The technique uses arguments based on the L^2 norm.

The paper will be divided into three main parts. General existence principles will be obtained in § 2. These will reduce our analysis to obtaining a priori bounds for a certain class of problems. In § 3 the L^2 theory will be extended to discuss the general boundary value problem (1.1). We use the L^n norm to obtain a priori bounds as described in § 2. In § 4 the Bernstein-Nagumo theory will be extended for problems of the above form.

To conclude this introduction we will use the following two theorems, which appear in the literature. Our first is a fixed point result, called the nonlinear alternative of Leray-Schauder [5], [7], [11]. By a map being *compact* we mean it is continuous with relatively compact range. A map is *completely continuous* if it is continuous and the image of every bounded set in the domain is contained in a compact subset of the range.

THEOREM 1.1. *Assume U is a relatively open subset of a convex set K in a Banach space E. Let $G: \bar{U} \to K$ be a compact map, $p \in U$, and $N_\lambda(u) = N(u, \lambda)$: $\bar{U} \times [0, 1] \to K$ a family of compact maps (i.e., $N(\bar{U} \times [0, 1])$ is contained in a compact subset of K and N: $\bar{U} \times [0, 1] \to K$ is continuous) with $N_1 = G$ and $N_0 = p$, the constant map to p. Then either*

(i) *G has a fixed point in $\bar{U}$; or*

(ii) *There is a point $u \in \partial U$ and $\lambda \in (0, 1)$ such that $u = N_\lambda u$.*

The next theorem involves results in integral inequalities [2], [3].

THEOREM 1.2. (a) *Suppose $y \in C^1[0, 1]$ with $y(0) = 0$. Then for any $p > 1$,*

$$\int_0^1 |y|^p \, dx \leqq \frac{1}{\mu_p} \int_0^1 |y'|^p \, dx,$$

where

$$\mu_p = \left(\frac{1}{p}(p-1)^{1/p} B\left(\frac{1}{p}, \frac{p-1}{p}\right)\right)^p = (p-1)\left(\int_0^1 \frac{ds}{(1-s^p)^{1/p}}\right)^p.$$

(b) *Suppose $y \in C^1[0, 1]$ with $y(0) = 0$. Then for any $p > 0$,*

$$\int_0^1 |y| \, |y'|^p \, dx \leqq \frac{1}{\eta_p} \int_0^1 |y'|^{p+1} \, dx,$$

where

$$\eta_p = 1 + \frac{2p \ln(p)}{p^2 - 1} \quad \textit{if } p \neq 1, \quad \textit{whereas } \eta_1 = 2.$$

(c) *Suppose $y \in C^1[0, 1]$ with $y(0) = y(1) = 0$. Then for any $p > 1$,*

$$\int_0^1 |y|^p \, dx \leqq \frac{1}{\sigma_p} \int_0^1 |y'|^p \, dx,$$

where

$$\sigma_p = 2^p(p-1)\left(\int_0^1 \frac{ds}{(1-s^p)^{1/p}}\right)^p.$$

(d) *Suppose* $y \in C^1[0,1]$ *with* $y(0)=y(1)=0$. *Then for any* $p>0$,

$$\int_0^1 |y||y'|^p\,dx \leqq \frac{1}{\omega_p}\int_0^1 |y'|^{p+1}\,dx,$$

where

$$\omega_p = 2\left[1+\frac{2p\ln(p)}{p^2-1}\right] \quad \textit{if } p\neq 1, \quad \textit{whereas } \eta_1 = 4.$$

Remarks. (i) When $p=2$, $\sigma_2=\pi^2$ and we have Wirtinger's inequality.

(ii) Note σ_p is the first eigenvalue [4] of

$$(\phi_p(y'))' + \lambda\phi_p(y) = 0, \quad 0<t<1, \quad \phi_p(u) = |u|^{p-2}u,$$
$$y(0)=y(1)=0.$$

(iii) When $p=1$, $\eta_1 = 2$ and we have Opial's inequality.

Proof. (a) and (b). These follow from [2, pp. 376-377].

(c) and (d). Since $y(0)=0$, a slight modification of the arguments in [2, pp. 376–379] yields for $p>0$, $q\geqq 0$ constants,

$$\int_0^{1/2} |y|^p|y'|^q\,dx \leqq K(p,q,p+q)\int_0^{1/2} |y'|^{p+q}\,dx, \tag{1.2}$$

where

$$K(p,q,p+q) = \frac{p(p/2)^p}{(p+q-1)(p+q)}(I(p,q,p+q))^{-p}$$

and

$$I(p,q,p+q) = \int_0^1 \left(1+\frac{(p+q)(q-1)t}{p}\right)^{(-1-p)/p}[1+(q-1)t]t^{(1/p)-1}\,dt.$$

In addition, since $y(1)=0$, a quick calculation yields

$$\int_{1/2}^1 |y|^p|y'|^q\,dx \leqq K(p,q,p+q)\int_{1/2}^1 |y'|^{p+q}\,dx. \tag{1.3}$$

Now (1.2) and (1.3) yield

$$\int_0^1 |y|^p|y'|^q\,dx \leqq K(p,q,p+q)\int_0^1 |y'|^{p+q}\,dx.$$

Notice that

$$K(p,0,p) = \frac{(p/2)^p}{(p-1)}\left(B\left(\frac{1}{p},\frac{p-1}{p}\right)\right)^{-p} = \frac{1}{2^p(p-1)}\left(\int_0^1 \frac{ds}{(1-s^p)^{1/p}}\right)^{-p},$$

which yields (c). Also notice that

$$K(1,q,q+1) = \frac{1}{2}\left(1+\frac{2q\ln(q)}{q^2-1}\right)^{-1} \quad \text{if } q\neq 1,$$

whereas $K(1,1,2)=\frac{1}{4}$, which yields (d). □

2. Existence principles. This section establishes existence principles for the general boundary value problem

$$(2.1)\qquad \begin{gathered}(\phi(y'))' = q(t)f(t, y, y'), \qquad 0<t<1,\\ y(0)=a, \qquad y(1)=b \;\text{ or }\; y'(0)=a, \qquad y(1)=b,\end{gathered}$$

where $q \in C(0, 1)$, f: $[0, 1]\times\mathbf{R}^2\to\mathbf{R}$ and $\phi:\mathbf{R}\to\mathbf{R}$ are continuous. The conditions on ϕ will be motivated by the physically reasonable situation [1],

$$\phi_n(u)=|u|^{n-2}u, \qquad 1<n\leqq 2.$$

We begin by obtaining an existence principle for the problem

$$(2.2)\qquad \begin{gathered}(\phi(y'))' = q(t)f(t, y, y'), \qquad 0<t<1,\\ y'(0)=a, \qquad y(1)=b.\end{gathered}$$

THEOREM 2.1. *Let $f:[0, 1]\times\mathbf{R}^2\to\mathbf{R}$ be continuous, and suppose*

$$(2.3)\qquad q\in C(0,1) \quad \textit{with } q>0 \quad \textit{on } (0,1)$$

$$(2.4)\qquad \begin{cases}\phi \textit{ is a continuous and strictly increasing map from}\\ (-\infty,\infty) \textit{ onto } (-\infty,\infty),\end{cases}$$

$$(2.5)\qquad \int_0^1 q(s)\,ds<\infty,$$

and

$$(2.6)\qquad \phi^{-1} \quad \textit{is continuously differentiable on } (-\infty,\infty)$$

are satisfied. In addition, assume there is a constant K^*, independent of λ, such that

$$|y|_1=\max\left\{\sup_{[0,1]}|y(t)|, \sup_{[0,1]}|y'(t)|\right\}\equiv\max\{|y|_0, |y'|_0\}\leqq K^*$$

for any solution y to

$$(2.7)_\lambda\qquad \begin{gathered}(\phi(y'))' = \lambda q(t)f(t, y, y'), \qquad 0<t<1,\\ y'(0)=a, \qquad y(1)=b\end{gathered}$$

for each $\lambda\in(0, 1)$. Then (2.2) *has at least one solution $y\in C^1[0, 1]$ with $(\phi(y'))'\in C(0, 1)$.*

Remark. When $\phi(u)=\phi_n(u)=|u|^{n-2}u$, then assumption (2.6) will be true if $1<n\leqq 2$.

Proof. Solving $(2.7)_\lambda$ is equivalent to finding a $y\in C^1[0, 1]$ that satisfies

$$(2.8)\qquad y(t)=b-\int_t^1\phi^{-1}\left(\phi(a)+\lambda\int_0^s q(u)f(u, y(u), y'(u))\,du\right)ds.$$

Define the operator $N_\lambda: C_B^1[0, 1]\to C_B^1[0, 1]$ by

$$(N_\lambda y)(t)=b-\int_t^1\phi^{-1}\left(\phi(a)+\lambda\int_0^s q(u)f(u, y(u), y'(u))\,du\right)ds.$$

Here $C_B^1[0, 1]=\{u\in C^1[0, 1]$: $u'(0)=a$, $u(1)=b\}$. Of course $(2.7)_\lambda$ is equivalent to the fixed point problem $y=N_\lambda y$. Certainly N_λ is continuous since ϕ^{-1} is and completely continuous by the Arzela-Ascoli theorem. To see this let $\Omega\subseteq C_B^1[0, 1]$ be bounded, i.e., $|u|_1 \leqq M$ for all $u \in \Omega$; here $M > 0$ is a constant. Firstly $N_\lambda \Omega$ is bounded. This follows from the inequalities

$$N_\lambda u(t)| \leqq |b| + \int_0^1 G(s)\, ds \quad \text{and} \quad |(N_\lambda u)'(t)| \leqq G(1),$$

where

$$G(s) = \max\left\{\left|\phi^{-1}\left(-|\phi(a)| - M_1 \int_0^s q(u)\, du\right)\right|, \left|\phi^{-1}\left(|\phi(a)| + M_1 \int_0^s q(u)\, du\right)\right|\right\} \tag{2.9}$$

and $M_1 = \sup |f(t, w_1, w_2)|$, where the supremum is computed over $[0, 1] \times [-M, M] \times [-M, M]$. We next show the equicontinuity of $N_\lambda \Omega$ on $[0, 1]$. For $u \in \Omega$ and $t, s \in [0, 1]$ we have

$$\begin{aligned} |N_\lambda u(t) - N_\lambda u(s)| &= \left|\int_s^t \phi^{-1}\left(\phi(a) + \lambda \int_0^v q(z) f(z, u(z), u'(z))\, dz\right) dv\right| \\ &\leqq \left|\int_s^t G(v)\, dv\right|, \end{aligned} \tag{2.10}$$

where G is described in (2.9). In addition, the Mean Value theorem yields

$$\begin{aligned} &|(N_\lambda u)'(t) - (N_\lambda u)'(s)| \\ &= \left|\phi^{-1}\left(\phi(a) + \lambda \int_0^t q(z) f(z, u(z), u'(z))\, dz\right)\right. \\ &\qquad \left. - \phi^{-1}\left(\phi(a) + \lambda \int_0^s q(z) f(z, u(z), u'(z))\, dz\right)\right| \\ &\leqq \sup |(\phi^{-1})'(x)| \left|\int_s^t q(z) f(z, u(z), u'(z))\, dz\right| \\ &\leqq M_1 \sup |(\phi^{-1})'(x)| \left|\int_s^t q(z)\, dz\right|, \end{aligned} \tag{2.11}$$

where the supremum is computed over $[-|\phi(a)| - M_1 \int_0^1 q(z)\, dz,\ |\phi(a)| + M_1 \int_0^1 q(z)\, dz]$, and $M_1 = \sup f(t, w_1, w_2)|$ with the supremum computed over $[0, 1] \times [-M, M] \times [-M, M]$. The equicontinuity of $N_\lambda \Omega$ on $[0, 1]$ now follows from (2.10), (2.11), and assumptions (2.5) and (2.6). Thus the Arzela-Ascoli theorem implies that N_λ is completely continuous. Set

$$U = \{u \in C_B^1[0, 1] : |u|_1 < K^* + 1\}, \qquad K = C_B^1[0, 1],$$
$$E = C^1[0, 1], \qquad N_0 y(t) = at + (b - a).$$

Remark. Note that $N(\bar{U} \times [0, 1])$ is contained in a compact subset of K. To see this let $N(u_n, \lambda_n)$ be any sequence in $N(\bar{U} \times [0, 1])$. Then as above $N(u_n, \lambda_n)$ is uniformly bounded (since G is (2.9) does not depend on the λ chosen) and equicontinuous on $[0, 1]$; so the Azela-Ascoli theorem again yields the result.

Apply Theorem 1.1 to deduce that N_1 has a fixed point, i.e., (2.2) has a solution $y \in C^1[0, 1]$. The fact that $(\phi(y'))' \in C(0, 1)$ follows from (2.8) with $\lambda = 1$. □

It is possible to improve the results of Theorem 2.1 if f is independent of its third variable, i,e., consider the boundary value problem

$$(2.12)\qquad \begin{gathered}(\phi(y'))'=q(t)f(t,y),\qquad 0<t<1,\\ y'(0)=a,\qquad y(1)=b.\end{gathered}$$

THEOREM 2.2. *Suppose* $f\colon[0,1]\times\mathbf{R}\to\mathbf{R}$ *is continuous and* (2.3), (2.4), *and* (2.5) *are satisfied. In addition, assume there is a constant* K^*, *independent of* λ, *such that* $|y|_0\leqq K^*$ *for any solution* y *to*

$$(2.13)_\lambda\qquad \begin{gathered}(\phi(y'))'=\lambda q(t)f(t,y),\quad 0<t<1,\\ y'(0)=a,\qquad y(1)=b\end{gathered}$$

for each $\lambda\in(0,1)$. *Then* (2.12) *has at least one solution* $y\in C^1[0,1]$ *with* $(\phi(y'))'\in C(0,1)$.

Remark. When $\phi(u)=\phi_n(u)=|u|^{n-2}u$, then since assumption (2.6) need *not* be satisfied in this situation, it is sufficient that $n>1$.

Proof. Solving $(2.13)_\lambda$ is equivalent to finding a $y\in C[0,1]$ that satisfies

$$(2.14)\qquad y(t)=b-\int_t^1\phi^{-1}\left(\phi(a)+\lambda\int_0^s q(u)f(u,y(u))\,du\right)ds.$$

Define the operator $N_\lambda\colon C_B[0,1]\to C_B[0,1]$. Here $C_B[0,1]=\{u\in C[0,1]\colon u(1)=b\}$, by

$$(N_\lambda y)(t)=b-\int_t^1\phi^{-1}\left(\phi(a)+\lambda\int_0^s q(u)f(u,y(u))\,du\right)ds.$$

Note N_λ is continuous and completely continuous (see (2.10)). Set

$$U=\{u\in C_B[0,1]\colon|u|_0<K^*+1\},\quad K=C_B[0,1],\quad E=C[0,1],$$
$$N_0y(t)=at+(b-a).$$

Now $N(\bar{U}\times[0,1])$ is contained in a compact subset of K; so apply Theorem 1.1 to deduce that N_1 has a fixed point y. The fact that $y\in C^1[0,1]$ with $y'(0)=a$ and $(\phi(y'))'\in C(0,1)$ follows from (2.14) with $\lambda=1$. □

We next obtain an existence principle for the problem

$$(2.15)\qquad \begin{gathered}(\phi(y'))'=q(t)f(t,y,y'),\quad 0<t<1,\\ y(0)=a,\qquad y(1)=b.\end{gathered}$$

THEOREM 2.3. *Suppose* $f\colon[0,1]\times\mathbf{R}^2\to\mathbf{R}$ *is continuous and* (2.3), (2.4), (2.5), *and* (2.6) *are satisfied. In addition, assume there is a constant* K^*, *independent of* λ, *such that* $|y|_1\leqq K^*$ *for any solution* y *to*

$$(2.16)_\lambda\qquad \begin{gathered}(\phi(y'))'=\lambda q(t)f(t,y,y'),\quad 0<t<1,\\ y(0)=a,\qquad y(1)=b\end{gathered}$$

for each $\lambda\in(0,1)$. *Then* (2.15) *has at least one solution* $y\in C^1[0,1]$ *with* $(\phi(y'))'\in C(0,1)$.

Proof. Solving $(2.16)_\lambda$ is equivalent to finding a $y\in C^1[0,1]$ which satisfies

$$(2.17)\qquad y(t)=a+\int_0^t\phi^{-1}\left(A-\lambda\int_s^1 q(z)f(z,y(z),y'(z))\,dz\right)ds,$$

where A satisfies

$$(2.18)\qquad b-a=\int_0^1\phi^{-1}\left(A-\lambda\int_s^1 q(z)f(z,y(z),y'(z))\,dz\right)ds.$$

We first show that A exists and is unique. Let $y \in C^1[0,1]$ be fixed with

$$H(x)=\int_0^1 \phi^{-1}(x-\lambda I(s))\,ds, \quad \text{where } I(s)=\int_s^1 q(z)f(z,y(z),y'(z))\,dz.$$

Certainly H is continuous since ϕ^{-1} is continuous (in particular, ϕ^{-1} is uniformly continuous on compact sets). Also there exists x_0, $x_1 \in [0,1]$ with $I(x_0) \leqq I(s) \leqq I(x_1)$ for all $s \in [0,1]$. Consequently, we have

$$\phi^{-1}(x-\lambda I(x_1)) \leqq H(x) \leqq \phi^{-1}(x-\lambda I(x_0)).$$

From this inequality it follows that H is a continuous function from $(-\infty,\infty)$ onto $(-\infty,\infty)$. Thus the Intermediate Value theorem implies there exists $A \in (-\infty,\infty)$ with $H(A)=b-a$. In addition, if $A_1 < A_2$, then since ϕ^{-1} is strictly increasing we have $H(A_1) < H(A_2)$. Thus A is unique.

Define the operator N_λ: $C_B^1[0,1] \to C_B^1[0,1]$. Here $C_B^1[0,1] = \{u \in C^1[0,1] : u(0)=a, u(1)=b\}$, by

$$(N_\lambda y)(t)=a+\int_0^t \phi^{-1}\left(A-\lambda\int_s^1 q(z)f(z,y(z),y'(z))\,dz\right)ds,$$

where A satisfies (2.18). We claim that $N_\lambda : C_B^1[0,1] \to C_B^1[0,1]$ is continuous. Let $u_n \to u$, $u_n' \to u'$ uniformly on $[0,1]$. We need to show that $N_\lambda u_n \to N_\lambda u$, $(N_\lambda u_n)' \to (N_\lambda u)'$ uniformly on $[0,1]$. Associate A_n with u_n and A with u in (2.18). Then

$$\begin{aligned} N_\lambda u_n(t)-N_\lambda u(t) = \int_0^t \Bigg(&\phi^{-1}\left(A_n-\lambda\int_s^1 q(z)f(z,u_n(z),u_n'(z))\,dz\right) \\ &-\phi^{-1}\left(A-\lambda\int_s^1 q(z)f(z,u(z),u'(z))\,dz\right)\Bigg)ds \end{aligned} \tag{2.19}$$

and

$$\begin{aligned} (N_\lambda u_n)'(t)-(N_\lambda u)'(t) = &\phi^{-1}\left(A_n-\lambda\int_t^1 q(z)f(z,u_n(z),u_n'(z))\,dz\right) \\ &-\phi^{-1}\left(A-\lambda\int_t^1 q(z)f(z,u(z),u'(z))\,dz\right), \end{aligned} \tag{2.20}$$

where

$$\begin{aligned} 0 = \int_0^1 \Bigg(&\phi^{-1}\left(A_n-\lambda\int_s^1 q(z)f(z,u_n(z),u_n'(z))\,dz\right) \\ &-\phi^{-1}\left(A-\lambda\int_s^1 q(z)f(z,u(z),u'(z))\,dz\right)\Bigg)ds. \end{aligned} \tag{2.21}$$

If we show $\lim_{n\to\infty} A_n = A$, then (2.19), (2.20), together with the fact that ϕ^{-1} is continuous, implies $N_\lambda : C_B^1[0,1] \to C_B^1[0,1]$ is continuous.

To see that $\lim_{n\to\infty} A_n = A$, notice that (2.21), together with the Mean Value theorem for integrals, implies that there exists $\eta_n \in [0,1]$ with

$$\begin{aligned} 0 = &\phi^{-1}\left(A_n-\lambda\int_{\eta_n}^1 q(z)f(z,u_n(z),u_n'(z))\,dz\right) \\ &-\phi^{-1}\left(A-\lambda\int_{\eta_n}^1 q(z)f(z,u(z),u'(z))\,dz\right). \end{aligned}$$

Thus

$$A_n - \lambda \int_{\eta_n}^{1} q(z)f(z, u_n(z), u_n'(z))\,dz = A - \lambda \int_{\eta_n}^{1} q(z)f(z, u(z), u'(z))\,dz,$$

and since $u_n \to u$, $u_n' \to u'$ uniformly on $[0, 1]$ we have $\lim_{n\to\infty} A_n = A$.

We next claim that N_λ is completely continuous. To see this let $\Omega \subseteq C_B^1[0, 1]$ be bounded, i.e., $|u|_1 \leqq M$, for all $u \in \Omega$. Here $M > 0$ is a constant. We first show that there exists a constant N^* with

$$|A_u| \leqq N^* \quad \text{for all } u \in \Omega.$$

Remark. Here A_u is given in (2.18), i.e., $A = A_u$ and $y = u$.

Since

$$b - a = \int_0^1 \phi^{-1}\left(A_u - \lambda \int_s^1 q(z)f(z, u(z), u'(z))\,dz\right) ds,$$

then the Mean Value theorem for integrals implies that there exists $\xi \in [0, 1]$ with

$$\phi^{-1}\left(A_u - \lambda \int_\xi^1 q(z)f(z, u(z), u'(z))\,dz\right) = b - a.$$

Consequently,

$$A_u = \lambda \int_\xi^1 q(z)f(z, u(z), u'(z))\,dz + \phi(b - a),$$

which implies

$$|A_u| \leqq M_1 \int_0^1 q(z) + \phi(b - a) \equiv N^*,$$

where $M_1 = \sup |f(t, w_1, w_2)|$ and the supremum is computed over $[0, 1] \times [-M, M] \times [-M, M]$.

Next we show that $N_\lambda\Omega$ is bounded. This follows from the following inequalities:

$$|N_\lambda u(t)| \leqq |a| + \int_0^1 J(s)\,ds \quad \text{and} \quad |(N_\lambda u)'(t)| \leqq J(0),$$

where

$$J(s) = \max\left\{\left|\phi^{-1}\left(-N^* - M_1 \int_s^1 q(u)\,du\right)\right|, \left|\phi^{-1}\left(N^* + M_1 \int_s^1 q(u)\,du\right)\right|\right\}.$$

We next show the equicontinuity of $N_\lambda\Omega$ on $[0, 1]$. For $u \in \Omega$ and $t, s \in [0, 1]$ we have (following the ideas of Theorem 2.1)

$$|N_\lambda u(t) - N_\lambda u(s)| \leqq \left|\int_s^t J(v)\,dv\right|$$

and

$$|(N_\lambda u)'(t) - (N_\lambda u)'(s)| \leqq M_1 \sup |(\phi^{-1})'(x)| \left|\int_s^t q(z)\,dz\right|,$$

where the supremum is computed over $[-N^*-M_1\int_0^1 q(z)\,dz, N^*+M_1\int_0^1 q(z)\,dz]$. Thus the Arzela-Ascoli theorem implies that N_λ is completely continuous. Set

$$U=\{u\in C_B^1[0,1]:|u|_1<K^*+1\},\ K=C_B^1[0,1],\ E=C^1[0,1],\ N_0y(t)=a+(b-a)t$$

and apply Theorem 1.1 to deduce the result. □

Again the result of Theorem 2.3 can be improved if we are interested in solutions to

$$(2.22)\qquad \begin{aligned}&(\phi(y'))'=q(t)f(t,y),\quad 0<t<1,\\ &y(0)=a,\qquad y(1)=b.\end{aligned}$$

THEOREM 2.4. *Suppose $f:[0,1]\times\mathbf{R}\to\mathbf{R}$ is continuous and* (2.3), (2.4), *and* (2.5) *are satisfied. In addition, assume there is a constant K^*, independent of λ, such that $|y|_0\leqq K^*$ for any solution y to*

$$(2.23)_\lambda\qquad \begin{aligned}&(\phi(y'))'=\lambda q(t)f(t,y),\quad 0<t<1,\\ &y(0)=a,\qquad y(1)=b\end{aligned}$$

for each $\lambda\in(0,1)$. Then (2.22) *has at least one solution $y\in C^1[0,1]$ with $(\phi(y'))'\in C(0,1)$.*

Proof. In this case the operator $N_\lambda: C_B[0,1]\to C_B[0,1]$, here $C_B[0,1]=\{u\in C[0,1]: u(0)=a,\ u(1)=b\}$, is defined by

$$(N_\lambda y)(t)=a+\int_0^t \phi^{-1}\left(A-\lambda\int_s^1 q(u)f(u,y(u))\,du\right)ds,$$

where

$$b-a=\int_0^1\phi^{-1}\left(A-\lambda\int_s^1 q(u)f(u,y(u))\,du\right)ds. \qquad □$$

3. A priori bounds using the $\mathbf{L}^n$-norm. The existence principles of the previous section are now used to examine (1.1) with Dirichlet or mixed boundary data. The cases where q is nonsingular or singular are discussed separately. We remark here that the results for the case $n=2$ extend and complement those in [12], [13]. The results reported in this section are for homogeneous boundary data; however, the nonhomogeneous case could be considered by a change of variables [13].

3.1. Mixed boundary data. In this subsection we discuss the boundary value problems

$$(3.1)\qquad \begin{aligned}&(\phi_n(y'))'=q(t)f(t,y,y'),\quad 0<t<1,\quad 1<n\leqq 2,\\ &y'(0)=y(1)=0,\end{aligned}$$

and

$$(3.2)\qquad \begin{aligned}&(\phi_n(y'))'=q(t)f(t,y),\quad 0<t<1,\quad n>1,\\ &y'(0)=y(1)=0,\end{aligned}$$

where $\phi_n(u)=|u|^{n-2}u$.

THEOREM 3.1 (nonsingular case). *Suppose $f:[0,1]\times\mathbf{R}^2\to\mathbf{R}$ is continuous and $q\in C[0,1]$ satisfies $q>0$ on $(0,1)$. In addition, assume f has the decomposition $f(t,u,v)=g(t,u,v)+h(t,u,v)$ with $g,h:[0,1]\times\mathbf{R}^2\to\mathbf{R}$ continuous. Now suppose*

$$(3.3)\quad |h(t,u,v)|\leqq K\{|u|^\alpha+|v|^\beta+1\}\ \textit{for constants } K,\alpha,\beta\ \textit{ with } 0\leqq\alpha,\beta<n-1,$$

for $a, b \in \mathbf{R}$, $ug(t, u, v) \geqq a|u|^n + b|u||v|^{n-1}$ for $(t, u, v) \in [0, 1] \times \mathbf{R}^2$ and

(3.4) $|g(t, u, v)| \leqq A(t, u)|v|^n + B(t, u)$, *where $A(t, u)$ and $B(t, u)$ are bounded on bounded sets.*

Then (3.1) *has a solution $y \in C^1[0, 1]$ with $(\phi_n(y'))' \in C[0, 1]$ in each of the following cases:*

(i) $a \geqq 0, b \geqq 0$;
(ii) $a < 0$, $b \geqq 0$ *and* $|a|N_0 < \mu_n$, *where* $N_0 = \sup_{[0,1]} q(t)$;
(iii) $a \geqq 0$, $b < 0$ *and* $|b|N_0 < \eta_{n-1}$;
(iv) $a < 0$, $b < 0$ *and* $|a|N_0\eta_{n-1} + |b|N_0\mu_n < \mu_n\eta_{n-1}$,

where η_{n-1} and μ_n are as described in Theorem 1.2.

Proof. To prove existence of a solution we apply Theorem 2.1. Let y be a solution to

$$(3.5)_\lambda \qquad \begin{aligned}&(\phi_n(y'))' = \lambda q(t)f(t, y, y'), \quad 0 < t < 1, \quad 1 < n \leqq 2, \quad 0 < \lambda < 1,\\ &y'(0) = y(1) = 0.\end{aligned}$$

Now $(\phi_n(y'))' = \lambda qg(t, y, y') + \lambda qh(t, y, y')$, together with $\int_0^1 (\phi_n(y'))' y\, dx = -\int_0^1 |y'|^n dx$, implies

$$(3.6) \qquad \int_0^1 |y'|^n dt \leqq -\lambda \int_0^1 qyg(t, y, y')\, dt + \int_0^1 q|y||h(t, y, y')|\, dt.$$

For notational purposes let

$$\|y_n\|_n = \left\{\int_0^1 |y|^n dt\right\}^{1/n}.$$

In addition, (3.3) yields

$$(3.7) \quad \int_0^1 q|y||h(t, y, y')|\, dt \leqq K \int_0^1 q|y|^{\alpha+1} dt + K \int_0^1 q|y||y'|^\beta dt + K \int_0^1 q|y|\, dt.$$

Now Hölder's inequality yields

$$(3.8) \qquad \sup_{[0,1]} |y(t)| = \sup_{[0,1]} \left|\int_0^t y'(s)\, ds\right| \leqq \|y'\|_n,$$

and so

$$(3.9) \qquad \int_0^1 q|y|^{\alpha+1} dt \leqq K_1\|y'\|_n^{\alpha+1} \quad \text{and} \quad \int_0^1 q|y|\, dt \leqq K_1\|y'\|_n,$$

where $K_1 = \int_0^1 q(s)\, ds$. Also Hölder's inequality implies

$$(3.10) \qquad \begin{aligned}\int_0^1 q|y||y'|^\beta dt &\leqq \|y'\|_n \int_0^1 q|y'|^\beta dt\\ &\leqq \|y'\|_n\left\{\int_0^1 [q(t)]^{n/(n-\beta)} dt\right\}^{(n-\beta)/n} \|y'\|_n^\beta = K_2\|y'\|_n^{\beta+1},\end{aligned}$$

where$K_2 = \{\int_0^1 [q(t)]^{n/(n-\beta)} dt\}^{(n-\beta)/n}$. Putting (3.9) and (3.10) into (3.7) and then (3.6) yields

$$(3.11) \quad \|y'\|_n^n \leqq -\lambda \int_0^1 qyg(t, y, y')\, dt + KK_1\|y'\|_n^{\alpha+1} + KK_2\|y'\|_n^{\beta+1} + KK_1\|y'\|_n.$$

Case (i). $a \geqq 0$, $b \geqq 0$.

Then $yg(t, y, y') \geqq 0$, and so (3.11) implies

$$\|y'\|_n^n \leqq KK_1\|y'\|_n^{\alpha+1} + KK_2\|y'\|_n^{\beta+1} + KK_1\|y'\|_n.$$

Thus there exists a constant M_0 independent of λ with

$$\|y'\|_n \leqq M_0 \tag{3.12}$$

for each solution y to $(3.5)_\lambda$.

Case (ii). $a<0$, $b \geqq 0$.

Then with $N_0 = \sup_{[0,1]} q(t)$,

$$-\lambda \int_0^1 qyg(t, y, y')\,dt \leqq (-a)\int_0^1 q|y|^n\,dt \leqq \frac{(-a)N_0}{\mu_n}\|y'\|_n^n \tag{3.13}$$

by Theorem 1.2. Putting (3.13) into (3.11) yields

$$\left(1+\frac{aN_0}{\mu_n}\right)\|y'\|_n^n \leqq KK_1\|y'\|_n^{\alpha+1} + KK_2\|y'\|_n^{\beta+1} + KK_1\|y'\|_n,$$

and we again have (3.12) since $|a|N_0 < \mu_n$.

Case (iii). $a \geqq 0$, $b<0$.

Then

$$-\lambda \int_0^1 qyg(t, y, y')\,dt \leqq (-b)N_0\int_0^1 |y|\,|y'|^{n-1}\,dt \leqq \frac{(-b)N_0}{\eta_{n-1}}\|y'\|_n^n \tag{3.14}$$

by Theorem 1.2. Putting (3.14) into (3.11) yields

$$\left(1+\frac{bN_0}{\eta_{n-1}}\right)\|y'\|_n^n \leqq KK_1\|y'\|_n^{\alpha+1} + KK_2\|y'\|_n^{\beta+1} + KK_1\|y'\|_n,$$

and we again have (3.12).

Case (iv). $a<0$, $b<0$.

Combining Cases (ii) and (iii) will again yield (3.12) in this case.

Now (3.8) and (3.12) yield

$$\sup_{[0,1]} |y(t)| \leqq \|y'\|_n \leqq M_0 \tag{3.15}$$

for any solution y to $(3.5)_\lambda$. This bound, together with (3.3) and (3.4), implies that there exist constants A and B independent of λ with

$$|(\phi_n(y'))'| \leqq Aq|y'|^n + Bq.$$

Consequently,

$$\int_0^1 |(\phi_n(y'))'|\,dt \leqq AN_0\|y'\|_n^n + BK_1 \leqq AN_0M_0^n + BK_1 \equiv M_2, \tag{3.16}$$

where $N_0 = \sup_{[0,1]} q(t)$ and $K_1 = \int_0^1 q(s)\,ds$.

In addition, since

$$|y'(t)|^{n-2}y'(t) = \int_0^t (|y'(s)|^{n-2}y'(s))'\,ds = \int_0^t |(\phi_n(y'(s)))'|\,ds,$$

we have for $t\in[0,1]$,

$$\||y'(t)|^{n-2}y'(t)|\leqq\int_0^1|(\phi_n(y'(s)))'|\,ds\leqq M_2$$

by (3.16). Consequently, $|y'(t)|\leqq M^{1/(n-1)}\equiv M_1$, and so

$$\sup_{[0,1]}|y'(t)|\leqq M_1 \tag{3.17}$$

for each solution y to $(3.5)_\lambda$.

Thus (3.15), (3.17), together with Theorem 2.1, yield the result. □

We also immediately have the following.

THEOREM 3.2. *Suppose* $f:[0,1]\times\mathbf{R}\to\mathbf{R}$ *is continuous and* $q\in C[0,1]$ *satisfies* $q>0$ *on* $(0,1)$. *In addition, assume* f *has the decomposition* $f(t,u)=g(t,u)+h(t,u)$ *with* $g,h:[0,1]\times\mathbf{R}\to\mathbf{R}$ *continuous. Now suppose*

$$|h(t,u)|\leqq K\{|u|^\alpha+1\}\quad \textit{for constants } K,\alpha\quad \textit{with } 0\leqq\alpha<n-1, \tag{3.18}$$

$$\textit{for } a\in\mathbf{R},\quad ug(t,u)\geqq a|u|^n\quad \textit{for } (t,u)\in[0,1]\times\mathbf{R}. \tag{3.19}$$

Then (3.2) *has a solution* $y\in C^1[0,1]$ *with* $(\phi_n(y'))'\in C(0,1)$ *in each of the following cases*:

(i) $a\geqq0$;

(ii) $a<0$ *and* $|a|N_0<\mu_n$, *where* $N_0=\sup_{[0,1]}q(t)$ *and* μ_n *is as described in Theorem* 1.2.

Proof. This follows from the ideas of Theorem 3.1 except we now use Theorem 2.2. □

THEOREM 3.3 (singular case). *Suppose* $f:[0,1]\times\mathbf{R}^2\to\mathbf{R}$ *is continuous with assumptions* (2.3) *and* (2.5) *being satisfied. In addition assume* f *has the decomposition* $f(t,u,v)=g(t,u,v)+h(t,u,v)$ *with* $g,h:[0,1]\times\mathbf{R}^2\to\mathbf{R}$ *continuous. Now suppose* (3.3) *holds and*

$$\begin{gathered}\textit{for } a,b\in\mathbf{R},\ ug(t,u,v)\geqq a|u|^n+b|u|\,|v|^{n-1}\ \textit{for } (t,u,v)\in[0,1]\times\mathbf{R}^2\ \textit{and}\\ \textit{there exists } \gamma,\ 0\leqq\gamma<n\ \textit{with } |g(t,u,v)|\leqq A(t,u)|v|^\gamma+B(t,u),\\ \textit{where } A(t,u)\ \textit{and } B(t,u)\ \textit{are bounded on bounded sets}\end{gathered} \tag{3.20}$$

$$\int_0^1 q^\theta(s)\,ds<\infty,\quad \textit{where } \theta=\max\left\{\frac{n}{n-\gamma},\frac{n}{n-\beta}\right\}. \tag{3.21}$$

Then (3.1) *has a solution* $y\in C^1[0,1]$ *with* $(\phi_n(y'))'\in C(0,1)$ *in each of the following cases*:

(i) $a\geqq0$, $b\geqq0$;

(ii) $a<0$, $b\geqq0$ *and* $|a|Q_0<1$, *where* $Q_0=\int_0^1 q(t)\,dt$;

(iii) $a\geqq0$, $b<0$, $Q_1<\infty$ *and* $|b|Q_1<1$, *where* $Q_1=(\int_0^1 q^n(s)\,ds)^{1/n}$;

(iv) $a<0$, $b<0$, $Q_1<\infty$ *and* $|a|Q_0+|b|Q_1<1$.

Proof. Exactly the same reasoning as in Theorem 3.1 yields

$$\|y'\|_n^n\leqq-\lambda\int_0^1 qyg(t,y,y')\,dt+KK_1\|y'\|_n^{\alpha+1}+KK_2\|y'\|_n^{\beta+1}+KK_1\|y'\|_n \tag{3.22}$$

for any solution y to $(3.5)_\lambda$.

Case (i). $a \geqq 0$, $b \geqq 0$.
Then

$$\|y'\|_n^n \leqq KK_1\|y'\|_n^{\alpha+1} + KK_2\|y'\|_n^{\beta+1} + KK_1\|y'\|_n.$$

So there exists a constant M_0 independent of λ with

$$\|y'\|_n \leqq M_0. \tag{3.23}$$

Case (ii). $a<0$, $b \geqq 0$.
Then

$$-\lambda \int_0^1 qyg(t, y, y')\,dt \leqq (-a)\|y'\|_n^n \int_0^1 q(t)\,dt = (-a)Q_0\|y'\|_n^n \tag{3.24}$$

using (3.8). Putting (3.24) into (3.22) yields

$$(1+aQ_0)\|y'\|_n^n \leqq KK_1\|y'\|_n^{\alpha+1} + KK_2\|y'\|_n^{\beta+1} + KK_1\|y'\|_n;$$

so again (3.23) is satisfied.

Case (iii). $a \geqq 0$, $b<0$.
Then Hölder's inequality yields

$$-\lambda \int_0^1 qyg(t, y, y')\,dt \leqq (-b)\|y'\|_n \int_0^1 q|y'|^{n-1}\,dt \leqq (-b)\|y'\|_n^n \left(\int_0^1 q^n(s)\,ds\right)^{1/n}. \tag{3.25}$$

Putting (3.25) into (3.22) yields

$$(1+bQ_1)\|y'\|_n^n \leqq KK_1\|y'\|_n^{\alpha+1} + KK_2\|y'\|_n^{\beta+1} + KK_1\|y'\|_n;$$

so again (3.23) is satisfied.

Case (iv). $a<0$, $b<0$. Combining Cases (ii) and (iii) will again yield (3.23) in this case.

Thus

$$\sup_{[0,1]} |y(t)| \leqq \|y'\|_n \leqq M_0 \tag{3.26}$$

for any solution y to $(3.5)_\lambda$. In addition, there exist constants A and B independent of λ with

$$|(\phi_n(y'))'| \leqq Aq|y'|^\tau + Bq,$$

where $\tau = \max\{\beta, \gamma\}$. Now Hölder's inequality yields

$$\begin{aligned}\int_0^1 |(\phi_n(y'))'|\,dt &\leqq A\|y'\|_n^\tau \left(\int_0^1 [q(t)]^{n/(n-\tau)}\,dt\right)^{(n-\tau)/n} + B\int_0^1 q(t)\,dt \\ &\leqq AM_0^\tau \left(\int_0^1 [q(t)]^{n/(n-\tau)}\,dt\right)^{(n-\tau)/n} + B\int_0^1 q(t)\,dt \equiv M_2\end{aligned}$$

and the result now follows as in Theorem 3.1. □

THEOREM 3.4. *Suppose* $f:[0,1]\times\mathbf{R}\to\mathbf{R}$ *is continuous with assumptions* (2.3) *and* (2.5) *being satisfied. In addition, assume* f *has the decomposition* $f(t,u)=g(t,u)+h(t,u)$ *with* $g, h:[0,1]\times\mathbf{R}\to\mathbf{R}$ *continuous. Now assume* (3.18) *and* (3.19) *are satisfied. Then* (3.2) *has a solution* $y\in C^1[0,1]$ *with* $(\phi_n(y'))'\in C(0,1)$ *in each of the following cases*:

(i) $a\geqq 0$;

(ii) $a<0$ *and* $|a|Q_0<1$, *where* $Q_0=\int_0^1 q(t)\,dt$.

3.2. Dirichlet boundary data. In this subsection we discuss the boundary value problems

$$(3.27)\qquad \begin{aligned}&(\phi_n(y'))' = q(t)f(t, y, y'), \quad 0<t<1, \quad 1<n\leqq 2,\\ &y(0)=y(1)=0,\end{aligned}$$

and

$$(3.28)\qquad \begin{aligned}&(\phi_n(y'))' = q(t)f(t, y), \quad 0<t<1, \quad n>1,\\ &y(0)=y(1)=0.\end{aligned}$$

Essentially the same reasoning as in § 3.1 immediately yields the following.

THEOREM 3.5. *Suppose* $f:[0,1]\times\mathbf{R}^2\to\mathbf{R}$ *is continuous and* $q\in C[0,1]$ *satisfies* $q>0$ *on* $(0,1)$. *In addition, assume* f *has the decomposition* $f(t,u,v)=g(t,u,v)+h(t,u,v)$ *with* $g, h:[0,1]\times\mathbf{R}^2\to\mathbf{R}$ *continuous. Now suppose* (3.3) *and* (3.4) *hold. Then* (3.27) *has a solution* $y\in C^1[0,1]$ *with* $(\phi_n(y'))'\in C[0,1]$ *in each of the following cases*:

(i) $a\geqq 0$, $b\geqq 0$;

(ii) $a<0$, $b\geqq 0$ *and* $|a|N_0<\sigma_n$, *where* $N_0=\sup_{[0,1]} q(t)$;

(iii) $a\geqq 0$, $b<0$ *and* $|b|N_0<\omega_{n-1}$;

(iv) $a<0$, $b<0$ *and* $|a|N_0\omega_{n-1}+|b|N_0\sigma_n<\sigma_n\omega_{n-1}$, *where* ω_{n-1} *and* σ_n *are as described in Theorem* 1.2.

THEOREM 3.6. *Suppose* $f:[0,1]\times\mathbf{R}^2\to\mathbf{R}$ *is continuous with assumptions* (2.3) *and* (2.5) *being satisfied. In addition assume* f *has the decomposition* $f(t,u,v)=g(t,u,v)+h(t,u,v)$ *with* $g, h:[0,1]\times\mathbf{R}^2\to\mathbf{R}$ *continuous. Now suppose* (3.3), (3.20), *and* (3.21) *are satisfied. Then* (3.27) *has a solution* $y\in C^1[0,1]$ *with* $(\phi_n(y'))'\in C(0,1)$ *in each of the following cases*:

(i) $a\geqq 0$, $b\geqq 0$;

(ii) $a<0$, $b\geqq 0$, *and* $|a|Q_0<1$, *where* $Q_0=\int_0^1 q(t)\,dt$;

(iii) $a\geqq 0$, $b<0$, $Q_1<\infty$, *and* $|b|Q_1<1$, *where* $Q_1=(\int_0^1 q^n(s)\,ds)^{1/n}$;

(iv) $a<0, b<0, Q_1<\infty$, *and* $|a|Q_0+|b|Q_1<1$.

Remark. We also have obvious analogues of Theorems 3.2 and 3.4 for the Dirichlet boundary value problem (3.28).

Remark. If $n=2$ we can replace $ug(t,u,v)\geqq a|u|^2+b|u||v|$ in assumption (3.4) and (3.20) by $ug(t,u,v)\geqq a|u|^2+b|u||v|+cuv$, and the results of Theorems 3.5 and 3.6 are again true.

It is possible to improve the results of this section if Theorem 1.2 is replaced by inequalities of the form

$$\int_0^1 |y|^p q(x)\,dx\leqq C_{1,p}\int_0^1 |y'|^p\,dx \quad\text{and}\quad \int_0^1 |y||y'|^{p-1}q(x)\,dx\leqq C_{2,p}\int_0^1 |y'|^p\,dx,$$

where $C_{1,p}$ and $C_{2,p}$ are the best possible constants (see Theorem 3.8). Here $q\in C^1(0,1)$ is measurable and positive almost everywhere on [0, 1]. For completeness we state the extensions for the mixed boundary data. It should, however, be remarked that $C_{1,p}$ and $C_{2,p}$ are extremely difficult to compute exactly in many situations; so the following results are more theoretical than practical.

THEOREM 3.7. *Suppose* $f:[0,1]\times\mathbf{R}^2\to\mathbf{R}$ *is continuous and* f *has the decomposition* $f(t,u,v)=g(t,u,v)+h(t,u,v)$ *with* $g, h:[0,1]\times\mathbf{R}^2\to\mathbf{R}$ *continuous. Also suppose one of the following holds*:

(a) $q\in C[0,1]\cap C^1(0,1)$ *with* $q>0$ *on* $(0,1)$ *and assumptions* (3.3) *and* (3.4) *are satisfied*;

(b) $q\in C^1(0,1)$ *with assumptions* (2.3), (2.5), (3.3), (3.20), *and* (3.21) *are satisfied.*

Then (3.1) *has a solution in each of the following cases*:

(i) $a \geqq 0$, $b \geqq 0$;

(ii) $a < 0$, $b \geqq 0$ *and* $|a| C_{1,n} < 1$;

(iii) $a \geqq 0$, $b < 0$, $\int_0^1 q^n(s)\, ds < \infty$ *and* $|b| C_{2,n} < 1$;

(iv) $a < 0$, $b < 0$, $\int_0^1 q^n(s)\, ds < \infty$ *and* $|a| C_{1,n} + |b| C_{2,n} < 1$.

Proof. The only changes are that (3.10), (3.24), and (3.25) become

$$\int_0^1 |y|\,|y'|^\beta q(x)\, dx \leqq C_{2,\beta+1} \int_0^1 |y'|^{\beta+1}\, dx \leqq C_{2,\beta+1} \|y'\|_n^{\beta+1}, \tag{3.10)*}$$

$$(-a) \int_0^1 q(x)|y|^n\, dx \leqq (-a) C_{1,n} \|y'\|_n^n, \qquad a < 0, \tag{3.24)*}$$

and

$$(-b) \int_0^1 |y|\,|y'|^{n-1} q(x)\, dx \leqq (-b) C_{2,n} \|y'\|_n^n, \qquad b < 0. \qquad \square \tag{3.25)*}$$

Remark. There is an obvious result also for the boundary value problem (3.2).

The following theorem guarantees the existence of $C_{1,n}$ and $C_{2,n}$.

THEOREM 3.8. *Suppose* $q \in C^1(0, 1)$ *with assumptions* (2.3) *and* (2.5) *holding.*

(a) *Then for* $p > 1$,

$$\begin{aligned} &\lambda((u')^{p-1})' + q(x) u^{p-1} = 0, \\ &\lim_{x\to 0} u(x) = 0, \lim_{x\to 1} [u'(x)]^{p-1} = 0, \\ &\|u'\|_p = 1 \end{aligned} \tag{3.29}$$

has solutions (u, λ) *with* $u \in C^2(0, 1)$ *and* $u(x) > 0$, $u'(x) > 0$ *on* $(0, 1)$. *In addition, there is a largest value of* λ *such that* (3.29) *has a solution, and if* $C_{1,p}$ *denotes this value, then for any* $y \in C^1[0, 1]$,

$$\int_0^1 |y|^p q(x)\, dx \leqq C_{1,p} \int_0^1 |y'|^p\, dx.$$

(b) *Assume* $\int_0^1 q^p(x)\, dx < \infty$. *Then for* $p > 1$,

$$\begin{aligned} &(\lambda p(u')^{p-1} - (p-1) u (u')^{p-2} q(x))' + q(x)(u')^{p-1} = 0, \\ &\lim_{x\to 0} u(x) = 0, \lim_{x\to 1} (\lambda p(u')^{p-1} - (p-1) u (u')^{p-2} q(x)) = 0, \\ &\|u'\|_{p-1} = 1 \end{aligned} \tag{3.30}$$

has solutions (u, λ) *with* $u \in C^2(0, 1)$ *and* $u(x) > 0$, $u'(x) > 0$ *on* $(0, 1)$. *In addition, there is a largest value of* λ *such that* (3.30) *has a solution, and if* $C_{2,p}$ *denotes this value then for any* $y \in C^1[0, 1]$,

$$\int_0^1 |y|\,|y'|^{p-1} q(x)\, dx \leqq C_{2,p} \int_0^1 |y'|^p\, dx.$$

Proof. (a) This follows from [2, Thm. 1] once we show that $T: L^p \to L^p$ defined by

$$Tu(x) = [q(x)]^{1/p} \int_0^x u(t)\, dt$$

is compact.

(b) This follows from [2, Thm. 1] once we show that $G: L^p \to L^p$ defined by

$$Gu(x) = q(x) \int_0^x u(t)\, dt$$

is compact.

We will show that T is compact. A similar argument works for G. Recall [10] that if (T_n) is a sequence of compact linear operators from a normed space X into a Banach space Y and if $\|T_n - T\| \to 0$ as $n \to \infty$, then the limit operator T is compact.

Let T_n be defined by

$$T_n u(x) = [w_n(x)]^{1/p} \int_0^x u(t)\, dt,$$

where

$$w_n(x) = \begin{cases} q(x) & \text{if } q(x) \leqq n, \\ n & \text{if } q(x) > n. \end{cases}$$

We first show that $T_n: L^p \to C[0,1] \subseteq L^p$ is compact. To do this we apply the Arzela-Ascoli theorem. Fix $n \in \{1, 2, \dots\}$. Let $\Omega \subseteq L^p$ be bounded, i.e., there exist $M > 0$ with $\|u\|_p \leqq M$ for all $u \in \Omega$. Certainly $T_n\Omega$ is bounded and the equicontinuity on $[0,1]$ follows since for $t, s \in [0,1]$ and $u \in \Omega$, we have

$$\begin{aligned}|(T_n u)(t) - (T_n u)(s)| &= \left| [w_n(t)]^{1/p} \int_0^t u(z)\, dz - [w_n(s)]^{1/p} \int_0^s u(z)\, dz \right| \\ &\leqq [w_n(t)]^{1/p} \left| \int_s^t u(z)\, dz \right| + \left| \int_0^s u(z)\, dz \right| \left| [w_n(t)]^{1/p} - [w_n(s)]^{1/p} \right| \\ &\leqq [w_n(t)]^{1/p} M |t-s|^{(p-1)/p} + M \left| [w_n(t)]^{1/p} - [w_n(s)]^{1/p} \right|\end{aligned}$$

by Hölder's inequality. The equicontinuity of $T_n\Omega$ now follows from the continuity of w_n. Consequently, T_n is compact.

Remark. If q is continuous on $[0,1]$, then $T: L^p \to L^p$ is compact by the above argument with q replacing w_n.

We now show that $\|T_n - T\|_p \to 0$ as $n \to \infty$. Notice that

$$\begin{aligned}\|T_n(u) - T(u)\|_p &= \left(\int_0^1 |([w_n(x)]^{1/p} - [q(x)]^{1/p}) \int_0^x u(s)\, ds|^p\, dx \right)^{1/p} \\ &\leqq \|u\|_p \left(\int_0^1 |[w_n(x)]^{1/p} - [q(x)]^{1/p}|^p\, dx \right)^{1/p},\end{aligned}$$

and so by taking the supremum over all u of norm 1 yields

$$\|T_n - T\|_p \leqq \left(\int_0^1 |[w_n(x)]^{1/p} - [q(x)]^{1/p}|^p\, dx \right)^{1/p} \equiv c_n.$$

It remains to show $c_n \to 0$ as $n \to \infty$. If this is so, then T is compact. We first show that $\lim_{n\to\infty} \int_0^1 |w_n(x) - q(x)|\, dx = 0$. To do this notice that w_n is continuous on $[0,1]$ for each n and $\lim_{n\to\infty} w_n(x) = q(x)$ almost everywhere on $[0,1]$. Now since $w_n(x) \leqq q(x)$ almost everywhere on $[0,1]$, we may apply the Lebesgue dominated convergence theorem to deduce

$$\lim_{n\to\infty} \int_0^1 |w_n(x) - q(x)|\, dx = 0.$$

Finally, to show $\lim_{n\to\infty} c_n = \lim_{n\to\infty} \|w_n^{1/p} - q^{1/p}\|_p = 0$ we use the following result [14, p. 76]: *suppose* $f \in L^p$, $f_n \in L^p$ *with* $f_n \to f$ *almost everywhere on* $[0, 1]$ *and* $\lim_{n\to\infty} \|f_n\|_p = \|f\|_p$. *Then* $\lim_{n\to\infty} \|f_n - f\|_p = 0$.

With $f_n = w_n^{1/p}$, $f = q^{1/p}$ we see that

$$\lim_{n\to\infty} \|w_n^{1/p}\|_p = \lim_{n\to\infty} \left(\int_0^1 w_n(x)\,dx\right)^{1/p} = \left(\int_0^1 q(x)\,dx\right)^{1/p} = \|q^{1/p}\|_p,$$

and so the result follows. Consequently, $T: L^p \to L^p$ is compact. A similar argument shows that G is compact. □

4. Bernstein–Nagumo theory. This section extends the Bernstein-Nagumo theory [7] to boundary value problems of the form (1.1) with Dirichlet or mixed boundary data. We begin by examining

$$(4.1) \qquad \begin{aligned} &(\phi(y'))' = q(t)f(t, y), \quad 0 < t < 1, \\ &y(0) = a, \quad y(1) = b, \end{aligned}$$

and

$$(4.2) \qquad \begin{aligned} &(\phi(y'))' = q(t)f(t, y), \quad 0 < t < 1, \\ &y'(0) = a, \quad y(1) = b. \end{aligned}$$

THEOREM 4.1. (a) *Suppose* $f: [0, 1] \times \mathbf{R} \to \mathbf{R}$ *and suppose* (2.3), (2.4), *and* (2.5) *are satisfied. In addition, assume*

(4.3) *there is a constant* $M > 0$ *such that* $yf(t, y) > 0$ *for* $|y| > M$ *and for all* $t \in [0, 1]$.

Then (4.1) *has at least one solution* $y \in C^1[0, 1]$ *with* $(\phi(y'))' \in C(0, 1)$.

(b) *Suppose* $f: [0, 1] \times \mathbf{R} \to \mathbf{R}$, *and suppose* (2.3), (2.4), (2.5), *and* (4.3) *are satisfied. In addition, assume* $\phi(a) = 0$. *Then* (4.2) *has at least one soultion* $y \in C^1[0, 1]$ *with* $(\phi(y'))' \in C(0, 1)$.

Proof. (a) Suppose y is any solution to

$$(4.4)_\lambda \qquad \begin{aligned} &(\phi(y'))' = \lambda q(t)f(t, y), \quad 0 < t < 1, \quad 0 < \lambda < 1, \\ &y(0) = a, \quad y(1) = b. \end{aligned}$$

We claim $\sup_{[0,1]} |y(t)| \leqq \max\{M, |a|, |b|\} \equiv M_1$. Once this is shown the result follows from Theorem 2.4. If $|y|$ achieves its maximum at zero or 1, then, trivially $\sup_{[0,1]} |y(t)| \leqq M_1$. Suppose now that $|y|$ achieves a positive maximum at $t_0 \in (0, 1)$ with $|y(t_0)| > M$. Then $y'(t_0) = 0$. Now the differential equation yields

$$y(t_0)(\phi(y'))'(t_0) = \lambda q(t_0) y(t_0) f(t_0, y(t_0)) > 0.$$

Without loss of generality assume $y(t_0) > 0$. Then there exists a $\delta > 0$ with

$$(\phi(y'))'(t) > 0 \quad \text{for } t \in (t_0 - \delta, t_0 + \delta) \subseteq (0, 1).$$

Consequently,

$$(4.5) \qquad \phi(y'(t)) < \phi(0) \quad \text{for } t \in (t_0 - \delta, t_0),$$

whereas

$$(4.6) \qquad \phi(y'(t)) > \phi(0) \quad \text{for } t \in (t_0, t_0 + \delta).$$

Now if $\phi(0) \geqq 0$, then (4.6) implies $\phi(y'(t)) > 0$ for $t \in (t_0, t_0 + \delta)$, and since ϕ^{-1} is increasing we have $y'(t) > 0$ for $t \in (t_0, t_0 + \delta)$. This contradicts the maximality of $|y(t_0)| = y(t_0)$. On the other hand, if $\phi(0) < 0$, then (4.5) implies $\phi(y'(t)) < 0$ for $t \in (t_0 - \delta, t_0)$, and so $y'(t) < 0$ for $t \in (t_0 - \delta, t_0)$, a contradiction.

A similar contradiction is established if $y(t_0)<0$. Thus $|y(t_0)|\leqq M$, and our claim is established.

(b) Suppose y is any solution to

$$(4.7)_\lambda \qquad \begin{aligned} &(\phi(y'))' = \lambda q(t)f(t,y), \quad 0<t<1, \quad 0<\lambda<1, \\ &y'(0)=a, \quad y(1)=b. \end{aligned}$$

We claim $\sup_{[0,1]}|y(t)|\leqq \max\{M,|b|\}\equiv M_2$. If $|y|$ achieves its maximum at 1, then trivially $\sup_{[0,1]}|y(t)|\leqq M_2$. On the other hand, if $|y|$ achieves a nonzero maximum at $t_0\in(0,1)$, then as in part (a), we have $|y(t_0)|\leqq M$. Finally, suppose $|y|$ achieves a nonzero maximum at zero with $|y(0)|>M$. Then $y(0)f(0,y(0))>0$, and so this together with the differential equations yields

$$y(t)(\phi(y'))'(t)>0$$

for $t>0$ and close to zero. Assume without loss of generality that $y(0)>0$. Then there exists an interval $(0,\delta)$ with $(\phi(y'))'>0$ for $t\in(0,\delta)$. Integration from zero to $t(t<\delta)$ with the fact that $\phi(a)=0$ yields $\phi(y'(t))>0$ for $t\in(0,\delta)$. Thus $y'(t)>0$ for $t\in(0,\delta)$, which contradicts the maximality of $|y(0)|=y(0)$. A similar contradiction is established if $y(0)<0$. Thus $|y(0)|\leqq M$, and our claim is established. The result now follows from Theorem 2.2. □

Finally, in this paper we consider the more general problems

$$(4.8) \qquad \begin{aligned} &(\phi(y'))' = q(t)f(t,y,y'), \quad 0<t<1, \\ &y(0)=a, \quad y(1)=b, \end{aligned}$$

and

$$(4.9) \qquad \begin{aligned} &(\phi(y'))' = q(t)f(t,y,y'), \quad 0<t<1, \\ &y'(0)=0, \quad y(1)=b. \end{aligned}$$

THEOREM 4.2. *Suppose* $f:[0,1]\times\mathbf{R}^2\to\mathbf{R}$ *is continuous and* (2.3), (2.4), (2.5), *and* (2.6) *are satisfied. In addition, suppose*

(4.10) *there is a constant* $M\geqq 0$ *such that* $yf(t,y,0)>0$ *for* $|y|>M$ *and for all* $t\in[0,1]$,

(4.11) $\phi(-u)=-\phi(u)$ *for* $u\neq 0$,

and

(4.12) *there is a continuous function* $\psi:[0,\infty)\to(0,\infty)$ *such that* $|f(t,u,v)|\leqq\psi(|v|)$ *for* $(t,u)\in[0,1]\times[-M_3,M_3]$. *Here* $M_3=\max\{M,|a|,|b|\}$ *if we are examining* (4.8), *whereas* $M_3=\max\{M,|b|\}$ *if we are interested in* (4.9).

Also assume **one** *of the following holds*:

(4.13) q *is continuous on* $[0,1]$ *and*

$$2M_3\sup_{[0,1]}q(t)<\int_{\phi(c)}^{\infty}\frac{\phi^{-1}(u)}{\psi(\phi^{-1}(u))}\,du$$

or

$$(4.14) \qquad \int_0^1 q(t)\,dt<\int_{\phi(c)}^{\infty}\frac{du}{\psi(\phi^{-1}(u))}$$

or

$$\text{there exists } p>1 \text{ with } \int_0^1 [q(t)]^{p/(p-1)}\,dt<\infty \text{ and} \tag{4.15}$$
$$(2M_3)^{1/p}\left(\int_0^1 [q(t)]^{p/(p-1)}\,dt\right)^{(p-1)/p} < \int_{\phi(c)}^{\infty} \frac{(\phi^{-1}(u))^{1/p}}{\psi(\phi^{-1}(u))}\,du.$$

Here $c=|b-a|$ *if we are examining* (4.8), *whereas* $c=0$ *if we are interested in* (4.9). *Then* (4.8) *and* (4.9) *have at least one solution* $y\in C^1[0,1]$ *with* $(\phi(y'))'\in C(0,1)$.

Proof. Essentially the same reasoning as in Theorem 4.1 yields

$$\sup_{[0,1]}|y(t)|\leqq M_3 \tag{4.16}$$

for any solution y to

$$(\phi(y'))'=\lambda q(t)f(t,y,y'),\quad 0<t<1,\quad 0<\lambda<1, \qquad y(0)=a,\quad y(1)=b, \tag{4.17$_\lambda$}$$

or

$$(\phi(y'))'=\lambda q(t)f(t,y,y'),\quad 0<t<1,\quad 0<\lambda<1, \qquad y'(0)=0,\quad y(1)=b. \tag{4.18$_\lambda$}$$

Now each point $t\in[0,1]$ for which $|y'(t)|>c$ belongs to an interval $[\mu,\nu]$ such that y' maintains a fixed sign on $[\mu,\nu]$ and $|y'(\mu)|$ and/or $|y'(\nu)|$ is c.

Case (i). Suppose (4.13) holds.

To be definite assume $y'>0$ on (μ,ν) and $y'(\mu)=c$. Then $(\phi(y'))'\leqq q(t)\psi(y')$ for $t\in(\mu,\nu)$. Multiply by y' and integrate from μ to t to obtain

$$\int_\mu^t \frac{\phi^{-1}(\phi(y'))(\phi(y'))'}{\psi(\phi^{-1}(\phi(y')))}\,ds=\int_\mu^t \frac{y'(\phi(y'))'}{\psi(y')}\,ds\leqq \sup_{[0,1]} q(t)\int_0^t y'\,ds.$$

Thus

$$H(\phi(y'(t)))=\int_{\phi(c)}^{\phi(y'(t))}\frac{\phi^{-1}(u)}{\psi(\phi^{-1}(u))}\,du\leqq 2M_3\sup_{[0,1]} q(t),$$

and so since $H:[0,\infty)\to[0,\infty)$ is increasing we have

$$|y'(t)|=y'(t)\leqq \phi^{-1}\left(H^{-1}\left(2M_3\sup_{[0,1]} q(t)\right)\right)\equiv M_4.$$

The other possibilities are treated similarly, and the same bound on $|y'(t)|$ is obtained. Thus

$$\sup_{[0,1]}|y'(t)|\leqq \max\{M_4,|c|\}. \tag{4.19}$$

Now (4.16), (4.19), together with Theorems 2.1 or 2.3, yield the result.

Case (ii). Suppose (4.14) holds.

To be definite assume $y'<0$ on (μ,ν) and $y'(\mu)=-c$. Then $-(\phi(y'))'\leqq q(t)\psi(-y')$ for $t\in(\mu,\nu)$, and integrate from μ to t to obtain

$$\int_\mu^t \frac{-(\phi(y'))'}{\psi(\phi^{-1}(\phi(-y')))}\,ds\leqq\int_\mu^t q(s)\,ds.$$

Now using (4.11) we obtain

$$J(\phi(-y'(t)))=\int_{\phi(c)}^{\phi(-y'(t))}\frac{du}{\psi(\phi^{-1}(u))}\leqq\int_0^1 q(s)\,ds,$$

and so

$$|y'(t)|=-y'(t)\leqq\phi^{-1}\left(J^{-1}\left(\int_0^1 q(s)\,ds\right)\right)\equiv M_5.$$

The other possibilities are treated similarly, and the same bound on $|y'(t)|$ is obtained. Thus

$$\sup_{[0,1]}|y'(t)|\leqq\max\{M_5,|c|\}. \tag{4.20}$$

Now (4.16) and (4.20) yield the result.

Case (iii). Suppose (4.15) holds.

To be definite assume $y'>0$ on (μ,ν) and $y'(\mu)=c$. Then $(\phi(y'))'\leqq q(t)\psi(y')$ for $t\in(\mu,\nu)$. Multiply by $(y')^{1/p}$, and integrate from μ to t to obtain

$$\int_\mu^t\frac{[\phi^{-1}(\phi(y'))]^{1/p}(\phi(y'))'}{\psi(\phi^{-1}(\phi(y')))}\,ds\leqq\int_\mu^t q(s)(y'(s))^{1/p}\,ds.$$

Thus Hölder's inequality yields

$$G(\phi(y'(t)))=\int_{\phi(c)}^{\phi(y'(t))}\frac{[\phi^{-1}(u)]^{1/p}}{\psi(\phi^{-1}(u))}\,du\leqq(2M_3)^{1/p}\left(\int_0^1[q(t)]^{p/(p-1)}\,dt\right)^{(p-1)/p}.$$

Consequently,

$$|y'(t)|=y'(t)\leqq\phi^{-1}\left(G^{-1}\left((2M_3)^{1/p}\left(\int_0^1[q(t)]^{p/(p-1)}\,dt\right)^{(p-1)/p}\right)\right)\equiv M_6.$$

The other possibilities are treated similarly, and the same bound on $|y'(t)|$ is obtained. Thus

$$\sup_{[0,1]}|y'(t)|\leqq\max\{M_6,|c|\}. \qquad \square$$

REFERENCES

[1] L. E. Bobisud, *Steady state turbulent flow with reaction*, Rocky Mountain J. Math., 21 (1991), pp. 993-1007.

[2] D. Boyd, *Best constants in a class of integral inequalities*, Pacific J. Math., 30 (1969), pp. 367-383.

[3] J. Brink, *Inequalities involving* $\|f\|_p$ *and* $\|f^{(n)}\|_q$ *for f with n zeros*, Pacific J. Math., 42 (1972), pp. 289-311.

[4] M. del Pino, M. Elgueta, and R. Manasevich, *A homotopic deformation along p of a Leray-Schauder degree result and existence for* $(|u'|^{p-2}u')'+f(t,u)=0$, $u(0)=u(T)=0$, $p>1$, J. Differential Equations, 80 (1989), pp. 1-13.

[5] J. Dugundji and A. Granas, *Fixed Point Theory, Vol.* I, Monograf. Mat. PWN, Warsaw, 1982.

[6] J. R. Esteban and J. L. Vazquez, *On the equation of turbulent filtration in one-dimensional porous media*, Nonlinear Anal., 10 (1986), pp. 1303-1325.

[7] A. Granas, R. B. Guenther, and J. W. Lee, *Some general existence principles in the Caratheódory theory of nonlinear differential systems*, J. Math. Pures Appl., 70 (1991), pp. 153-196.

[8] M. A. Herrero and J. L. Vazquez, *On the propagation properties of a nonlinear degenerate parabolic equation*, Comm. Partial Differential Equations, 7 (1982), pp. 1381-1402.

[9] H. G. KAPER, M. KNAPP, AND M. K. KWONG, *Existence theorems for second order boundary value problems*, Differential and Integral Equations, 4 (1991), pp. 543-554.

[10] E. KREYSZIG, *Introductory Functional Analysis and Applications*, John Wiley, New York, 1978.

[11] J. W. LEE AND D. O'REGAN, *Existence results for differential delay equations* I, J. Differential Equations., to appear.

[12] D. O'REGAN, *Second and higher order systems of boundary value problems*, J. Math. Anal. Appl., 156 (1991), pp. 120-149.

[13] ———, *Boundary value problems for second and higher order differential equations*, Proc. Amer. Math. Soc., 113 (1991), pp. 761-776.

[14] W. RUDIN, *Real and Complex Analysis*, McGraw Hill, New York, 1966.

SIAM J. MATH. ANAL.
Vol. 24, No. 3, pp. 669-680, May 1993

008

NEGATIVE EIGENVALUES FOR A NONLINEAR DIFFERENTIAL EQUATION ON ℝ*

W. ROTHER†

Abstract. This paper considers a nonlinear eigenvalue problem involving a second-order ordinary differential equation on ℝ with variable coefficients. One of the coefficients may be negative on some subsets of ℝ and both may be unbounded. In case that the coefficients are positive constants this problem has been studied by H. Berestycki and P.-L. Lions [*Arch. Rational Mech. Anal.*, 82 (1983), pp. 313-345]. The paper shows that the problem has a negative eigenvalue and a positive classical solution decaying exponentially at infinity. Moreover, in some special cases the existence of bifurcation is proved. The main tools are direct methods from the calculus of variations, some comparison techniques, and Lebesgue's theorem on monotone functions.

Key words. nonlinear ordinary differential equation, negative eigenvalue, variational methods, bifurcation

AMS subject classification. 34B15

1. Introduction. In the present paper we study the nonlinear Dirichlet problem

$$(1.1)\qquad \begin{gathered} -u''-q(t)|u|^{\sigma_1}u+r(t)|u|^{\sigma_2}u=\lambda u \quad \text{for } t\in\mathbb{R}, \\ \lim_{t\to-\infty} u(t)=\lim_{t\to\infty} u(t)=0, \end{gathered}$$

where σ_1 and σ_2 are positive constants, λ is negative, and $q(\cdot)$ and $r(\cdot)$ are real-valued functions.

In case that $r\equiv 0$ and $N\geqq 2$ (replace u'' by Δu), this problem has been considered by many authors. See, for instance, [4], [5], [7]-[10], [13], [16]-[18], [20]-[23] and the literature quoted therein. In case $N=1$ and $r\equiv 0$, existence of nontrivial solutions and bifurcation in $L^p(\mathbb{R})$ has been proved in [19] and [20]. In [20] (respectively, [19]), the author assumes that the function q satisfies

$$0<A\leqq q(t)(1+|t|)^a\leqq B<\infty$$

for some constant $a\in[0,2)$ (respectively, $a=0$).

Moreover, Zhu and Zhou [23] proved that bifurcation occurs in $H^1(\mathbb{R})$ if $r\equiv 0$, $\sigma_1<4$, and q is a continuous function on $\mathbb{R}$ satisfying $q(t)\to q_\infty>0$ as $|t|\to\infty$ and $q(t)\geqq q_\infty$ for $t\in\mathbb{R}$.

Similar results for the Dirichlet problem on $(0,\infty)$ were obtained in [11], [15], and [17, Thm. 7.4].

In all cases quoted above the only bifurcation point is $\lambda=0$, i.e., the infimum of the essential spectrum of the linearization $-u''$ (respectively, $-\Delta u$).

In case q and r are positive constants and $N=1$ or $N\geqq 3$, existence results for (1.1) have been proved in [3] (see also [14, Thm. 2; Remark p. 160-161]).

Suppose, for instance, that $\sigma_1<\sigma_2$. Then, Berestycki and Lions (see [3, Ex. 2, Thms. 1 and 5]) have shown that (1.1) has a positive classical solution, decaying

* Received by the editors February 3, 1992; accepted for publication (in revised form) September 3, 1992.

† Department of Mathematics, University of Bayreuth, P.O.B. 101251, W-8580 Bayreuth, Germany.

exponentially at infinity if and only if $q > q^*$, where

$$q^* = \left(\frac{|\lambda|}{2}\right)^{(\sigma_2-\sigma_1)/\sigma_2} r^{\sigma_1/\sigma_2}\sigma_2(2+\sigma_1)(\sigma_2-\sigma_1)^{(\sigma_1-\sigma_2)/\sigma_2}\sigma_1^{-\sigma_1/\sigma_2}(2+\sigma_2)^{-\sigma_1/\sigma_2}.$$

Hence it seems to be an interesting question if (1.1) possesses a positive classical solution when the function q is negative on some subsets of $\mathbb{R}^N$ and $r(\cdot)$ is unbounded.

In [12] we presented some existence results for this problem when $N \geqq 3$, $0 < \sigma_1 < 4/(N-2)$, and $\sigma_2 \geqq 4/(N-2)$. We supposed that q and r satisfy some growth conditions on a certain subset $\mathscr{B}$ of $\mathbb{R}^N$, and outside of $\mathscr{B}$ we required the functions r and $q_- = \min(q, 0)$ to be locally integrable. To prove that the solutions decay exponentially at infinity we had to assume that the corresponding eigenvalues λ are negative (see Corollaries 1.1 and 1.2 in [12]), and had to show that some λ are negative. We had to impose rather restrictive conditions on the functions q and r (see [12, Thms. 1.3-1.5]). By the way, in case $r \equiv 0$, the eigenvalues are always negative, and the proof of this fact is evident (see, for instance, [7, p. 571] and Lemma 3.11).

In the following, we present some results for the case $N = 1$. It turns out that we need no additional assumptions in this case to show that $\lambda < 0$ holds for some eigenvalues λ (see Lemmas 3.12 and 4.3). Moreover, if $\sigma_1 < 2(2-a)$ (the constant a may be chosen as in condition (i.c)) we can show that there exists a sequence of solution pairs (u_n, λ_n), with $\lambda_n < 0$, bifurcating from $\lambda = 0$ (see Thm. 1.2). Using some standard techniques (see Lemma 3.10) and assuming that q is bounded above near infinity, we prove that the corresponding solutions u (respectively, u_n) decay exponentially at infinity. The main tool in the proof of Theorem 1.2 is Lebesgue's theorem on monotone functions. We use this theorem to show that the monotone decreasing energy $I(\cdot)$ is almost everywhere differentiable. Even in the case where $r \equiv 0$ the results of Theorems 1.1 and 1.2 seem to be new since q may change sign and may be unbounded below.

To present our results, we start with the formulation of some conditions for σ_1, σ_2, and the functions q and r. For the constants σ_1 and σ_2, we consider the cases

(i) $0 < \sigma_1 < 4$ and $0 < \sigma_2$; and

(ii) $4 \leqq \sigma_1$ and $\sigma_1 < \sigma_2$.

In case (i), we assume

(i.a) That the functions $q, r: \mathbb{R} \to \mathbb{R}$ are measurable, that r is nonnegative and that r and $q_- = \min(q, 0)$ are locally integrable;

(i.b) That the function $q_+ = \max(q, 0)$ can be written as $q_+ = q_1 + q_2$, where q_1 satisfies $0 \leqq q_1 \in L^\infty$ so that $q_1(t) \to 0$ as $|t| \to \infty$, and the function q_2 satisfies $0 \leqq q_2 \in L^{p_1}$ for some constant $p_1 \in [1, \infty) \cap (2/(4-\sigma_1), \infty)$;

(i.c) That there exist constants $t_0 \geqq 1$, $a > 0$, and $\mathscr{K} > 0$, and a measurable function $f: [t_0, \infty) \to [0, \infty)$, satisfying $f(t) \to \infty$ as $t \to \infty$, such that

(1) $q(t) \geqq f(t)t^{-a}$ and $r(t) \leqq \mathscr{K}t^b$ hold for all $t \geqq t_0$ or

(2) $q(t) \geqq f(-t)(-t)^{-a}$ and $r(t) \leqq \mathscr{K}(-t)^b$ hold for all $t \leqq -t_0$.

Here, the constant b is defined by $b = (2-a)(\sigma_2/\sigma_1) - 2$.

If (ii) is satisfied, we assume

(ii.a) That condition (i.a) holds true and that there exists a positive constant r_0 so that $r(t) \geqq r_0$ holds almost everywhere in $\mathbb{R}$;

(ii.b) that condition (i.b) is satisfied with the exception that the function q_2 satisfies $0 \leqq q_2 \in L^{p_2}$ for some $p_2 \in ((2+\sigma_2)/(\sigma_2-\sigma_1), \infty)$;

(ii.c) and that condition (i.c) holds true.

Remark 1.1. (a) The only assumption which the functions q_- and r have to fulfill on $(-\infty, t_0)$ (respectively, $(-t_0, \infty)$) is to be locally integrable. So q_- and r may have singularities on $(-\infty, t_0)$ (respectively, $(-t_0, \infty)$), q_- may decrease very fast to $-\infty$ as $t \to -\infty$ (respectively, $t \to \infty$) and r may increase very fast to $+\infty$ as $t \to -\infty$ (respectively, $t \to \infty$).

(b) In case (ii) it is assumed that $r(t) \geqq r_0 > 0$ holds for almost all $t \in \mathbb{R}$. Hence condition (ii.c) only can be fulfilled if $b \geqq 0$, i.e., if $a \leqq 2(1-(\sigma_1/\sigma_2))$.

(c) If v is a solution of the equation

$$-v'' - q(-t)|v|^{\sigma_1} v + r(-t)|v|^{\sigma_2} v = \lambda v \quad \text{in } \mathbb{R},$$

then $u(t) = v(-t)$ is a solution of (1.1). Hence, we may assume without restriction that part (1) of condition (i.c), respectively (ii.c), is satisfied.

Then we will prove the following results.

THEOREM 1.1. (a) *Suppose that* (i) *or* (ii) *holds true and that* $\sigma_1 \geqq 2(2-\mathrm{a})$. *Then, there exists a positive function* $u \in C^1 \cap H^1$ *and a negative constant* λ *such that* (1.1) *holds in the sense of distributions. In case* $\sigma_1 = 2(2-\mathrm{a})$, *we have* $\|u\|_2 = 1$.

(b) *If q and r are continuous, then u is twice continuously differentiable and* (1.1) *holds in the classical sense. Moreover, in case there exist positive constants* t_1 *and C such that*

$$q(t) \leqq C \quad \textit{holds for almost all } |t| \geqq t_1,$$

the function u decays exponentially as $|t| \to \infty$.

THEOREM 1.2. (a) *Suppose that* (i) *and* (ii) *hold true and that* $\sigma_1 < 2(2-\mathrm{a})$. *Then, there exist a sequence* (u_n) *of pairwise distinct functions* $u_n \in C^1 \cap H^1$ *and a sequence of negative constants* (λ_n) *such that* u_n *is positive and* (u_n, λ_n) *solves* (1.1) *in the sense of distributions. Letting* $n \to \infty$, *it follows that*

$$\|u_n\|_{H^1}^2 + \int |q||u_n|^{2+\sigma_1}\, dt + \int r|u_n|^{2+\sigma_2}\, dt \to 0$$

and $\lambda_n \to 0$. *In particular, we see that* $\lambda = 0$ *is a bifurcation point for* (1.1) *in* H^1 *and in* $L^2 \cup L^\infty$.

(b) *The assertions of Theorem* 1.1(b) *hold true if we replace u by* u_n.

2. Some preliminaries. By $L^p = L^p(\mathbb{R})$ and $L^p_{\text{loc}} = L^p_{\text{loc}}(\mathbb{R})$ $(1 \leqq p \leqq \infty)$, we denote the usual Lebesgue spaces, and $\|\cdot\|_p$ is the norm on L^p. If $1 < p < \infty$, the dual index p' is defined by $p' = p/(p-1)$. Moreover, H^1 denotes the Sobolev space $H^1(\mathbb{R}) = W^{1,2}(\mathbb{R})$, and the norm $\|\cdot\|_{H^1}$ is defined by $\|u\|_{H^1} = (\|u'\|_2^2 + \|u\|_2^2)^{1/2}$. Finally, $C^1 = C^1(\mathbb{R})$ is the space of the continuously differentiable functions, and $C_0^\infty = C_0^\infty(\mathbb{R})$ denotes the set of all functions that have compact support and derivatives of any order.

LEMMA 2.1. *Each function* $u \in H^1$ *can be identified with a Hölder continuous function on* $\mathbb{R}$, *still denoted by u, such that* $\lim_{|t| \to \infty} u(t) = 0$,

$$|u(t_1) - u(t_2)| \leqq \|u'\|_2 |t_1 - t_2|^{1/2} \quad \textit{holds for all } t_1, t_2 \in \mathbb{R},$$

and $\|u\|_\infty \leqq \sqrt{2}\, \|u'\|_2^{1/2} \|u\|_2^{1/2}$.

Proof. For $\varphi \in C_0^\infty$ we have

$$|\varphi(t_1) - \varphi(t_2)| = \left| \int_{t_2}^{t_1} \varphi'(s)\, ds \right| \leqq \|\varphi'\|_2 |t_1 - t_2|^{1/2}$$

and $\varphi^2(t) = 2\int_{-\infty}^t \varphi'(s)\varphi(s)\, ds$. Since C_0^∞ is dense in H^1, we obtain the assertion. □

3. Proof of Theorem 1.1. We start with the following lemma.

LEMMA 3.1. *Suppose that* (i) *and* (i.a)–(i.c) *hold true. Then there exist positive constants* α_1 *and* β_1, *and for each* $\varepsilon > 0$ *a constant* $C_{1,\varepsilon} \geqq 0$, *such that*

$$(2+\sigma_1)^{-1} \int q_+ |u|^{2+\sigma_1}\, dt \leqq \varepsilon \|u'\|_2^2 + C_{1,\varepsilon}(\|u\|_2^{2+\alpha_1} + \|u\|_2^{2+\beta_1})$$

holds for all $u \in H^1$.

Proof. From Lemma 2.1 and Hölder's inequality we conclude that

$$\int q_1 |u|^{2+\sigma_1}\, dt \leqq 2^{\sigma_1/2} \|q_1\|_\infty \|u'\|_2^{\sigma_1/2} \|u\|_2^{2+(\sigma_1/2)}.$$

The constant p_1 may be chosen as in condition (i.b). Then, in case $p_1 = 1$ we see that

$$\int q_2 |u|^{2+\sigma_1}\, dt \leqq 2^{1+(\sigma_1/2)} \|q_2\|_1 \|u'\|_2^{1+(\sigma_1/2)} \|u\|_2^{1+(\sigma_1/2)}.$$

Further, if $p_1 > 1$, we obtain

$$\begin{aligned}\int q_2 |u|^{2+\sigma_1}\, dt &\leqq \|q_2\|_{p_1} \|u\|_\infty^{2+\sigma_1-(2/p_1')} \|u\|_2^{2/p_1'} \\ &\leqq 2^{(1/p_1)+(\sigma_1/2)} \|q_2\|_{p_1} \|u'\|_2^{(1/p_1)+(\sigma_1/2)} \|u\|_2^{1+(1/p_1')+(\sigma_1/2)}.\end{aligned}$$

Since $\sigma_1 < 4$, $p_1 > 2/(4-\sigma_1)$, and $\sigma_1 < 2$ if $p_1 = 1$, the assertion follows from Young's inequality. □

LEMMA 3.2. *Suppose that* (ii) *and* (ii.a)–(ii.c) *hold true. Then there exist positive constants* α_2 *and* β_2, *and for each* $\varepsilon > 0$ *a constant* $C_{2,\varepsilon} \geqq 0$, *such that*

$$(2+\sigma_1)^{-1} \int q_+ |u|^{2+\sigma_1}\, dt \leqq \varepsilon \|u\|_{2+\sigma_2}^{2+\sigma_2} + C_{2,\varepsilon}(\|u\|_2^{\alpha_2} + \|u\|_2^{\beta_2})$$

holds for all $u \in H^1$.

Proof. For $\gamma_1 = \sigma_1/\sigma_2$, we see that $2+\sigma_1 = \gamma_1(2+\sigma_2) + (1-\gamma_1)2$. Hence, by Hölder's inequality, it follows that

$$\int q_1 |u|^{2+\sigma_1}\, dt \leqq \|q_1\|_\infty \left(\int |u|^{2+\sigma_2}\, dt \right)^{\gamma_1} \left(\int |u|^2\, dt \right)^{1-\gamma_1}.$$

The constant p_2 may be chosen as in condition (ii.b). Then it follows that $(2+\sigma_1)p_2' < 2+\sigma_2$. Thus, there exists a constant $\gamma_2 \in (0,1)$ such that $(2+\sigma_1)p_2' = \gamma_2(2+\sigma_2) + (1-\gamma_2)2$. Now, Hölder's inequality implies

$$\int q_2 |u|^{2+\sigma_1}\, dt \leqq \|q_2\|_{p_2} \left(\int |u|^{2+\sigma_2}\, dt \right)^{\gamma_2/p_2'} \left(\int |u|^2\, dt \right)^{(1-\gamma_2)/p_2'}.$$

Since $\gamma_1, \gamma_2/p_2' \in (0,1)$, the assertion follows by Young's inequality. □

In the following, we always assume that either (i) and (i.a)–(i.c) (part (1)) or (ii) and (ii.a)–(ii.c) (part (1)) are fulfilled. The functional ξ may be defined by

$$\xi(u) = \frac{1}{2} \int |u'|^2\, dt - (2+\sigma_1)^{-1} \int q(t)|u|^{2+\sigma_1}\, dt + (2+\sigma_2)^{-1} \int r(t)|u|^{2+\sigma_2}\, dt.$$

Moreover, for $\mu \geqq 0$, we define the set S_μ by

$$S_\mu = \left\{ u \in H^1;\ \int |q_-||u|^{2+\sigma_1}\, dt < \infty,\ \int r(t)|u|^{2+\sigma_2}\, dt < \infty \text{ and } \|u\|_2 \leqq \mu \right\}.$$

Since $r(t) \geqq r_0 > 0$ holds in case (ii), we conclude from Lemmas 3.1 and 3.2 that $\xi(\cdot)$ is bounded below on S_μ. Hence

$$I(\mu) = \inf_{u \in S_\mu} \xi(u)$$

is a well-defined real number.

LEMMA 3.3. *Let $k = (a-2)/\sigma_1$. Then there exists a function $h:[1,\infty) \to \mathbb{R}$ such that $h(s) \to -\infty$ as $s \to \infty$, and*

$$I(s^{k+(1/2)}) \leqq s^{2k-1} h(s) \quad \textit{holds for all } s \in [1,\infty).$$

Proof. The function $\varphi \in C_0^\infty$ may be chosen such that supp $\varphi \subset (t_0, \infty)$ and $\|\varphi\|_2 = 1$. Moreover, for $s \geqq 1$ we define the function φ_s by $\varphi_s(t) = s^k \varphi(s^{-1}t)$. Since $\|\varphi_s\|_2 = s^{k+(1/2)}$, we see that

$$\begin{aligned} I(s^{k+(1/2)}) \leqq \xi(\varphi_s) = s^{2k-1}\Big(\frac{1}{2}\int |\varphi'|^2\,dt - s^{2+\sigma_1 k}(2+\sigma_1)^{-1}\int_{t_0}^{\infty} q(st)|\varphi(t)|^{2+\sigma_1}\,dt \\ + s^{2+\sigma_2 k}(2+\sigma_2)^{-1}\int_{t_0}^{\infty} r(st)|\varphi(t)|^{2+\sigma_2}\,dt\Big). \end{aligned} \tag{3.1}$$

Now, from (i.c), respectively, (ii.c), we conclude that the right-hand side of (3.1) is less than or equal to $s^{2k-1}h(s)$, where

$$\begin{aligned} h(s) = \frac{1}{2}\int |\varphi'|^2\,dt - \inf_{x \geqq st_0} f(x)(2+\sigma_1)^{-1}\int_{t_0}^{\infty} t^{-a}|\varphi(t)|^{2+\sigma_1}\,dt \\ + \mathcal{K}(2+\sigma_2)^{-1}\int_{t_0}^{\infty} t^{b}|\varphi(t)|^{2+\sigma_2}\,dt. \end{aligned}$$

Since $\inf_{x \geqq st_0} f(x) \to \infty$ as $s \to \infty$, we obtain the assertion. □

LEMMA 3.4. *In case*

(i) $\sigma_1 > 2(2-a)$, *there exists a constant* $\mu > 0$ *such that* $I(\mu) < 0$;

(ii) $\sigma_1 = 2(2-a)$, *it follows that* $I(1) < 0$;

(iii) $\sigma_1 < 2(2-a)$, *there exists a constant* $\mu_0 > 0$ *such that* $I(\mu) < 0$ *holds for all* $\mu \in (0, \mu_0]$.

Proof. The constant k may be chosen as in Lemma 3.3. Then, in case (i) we have that $k+\frac{1}{2} > 0$, in case (ii) $k + \frac{1}{2} = 0$, and in case (iii) $k + \frac{1}{2} < 0$. Hence, the assertions follow from Lemma 3.3. □

LEMMA 3.5. *Suppose that there exists a $\mu > 0$ such that $I(\mu) < 0$. Moreover, the constants α_i and β_i $(i = 1, 2)$ may be chosen as in Lemma* 3.1, *respectively, Lemma* 3.2. *Then, there exists a nonnegative function $u \in S_\mu$ such that $\|u\|_2 > 0$ and $\xi(u) = I(\mu)$. Furthermore, there exists a constant C, independent of μ, so that*

$$\|u'\|_2^2 + \int |q_-||u|^{2+\sigma_1}\,dt + \int r|u|^{2+\sigma_2}\,dt \leqq C(\mu^{2+\alpha_1} + \mu^{2+\beta_1})$$

holds in case (i) *and*

$$\|u'\|_2^2 + \int |q_-||u|^{2+\sigma_1}\,dt + \int r|u|^{2+\sigma_2}\,dt \leqq C(\mu^{\alpha_2} + \mu^{\beta_2})$$

holds in case (ii).

Proof. We start with case (i). Let $(u_n) \subset S_\mu$ be a sequence such that $\xi(u_n) \to I(\mu)$. Then, we may assume without restriction that $\xi(u_n) \leqq 0$ and $u_n \geqq 0$. Hence, we obtain from Lemma 3.1 that

$$\begin{aligned}\frac{1}{4}\|u_n'\|_2^2 + (2+\sigma_1)^{-1}\int |q_-||u_n|^{2+\sigma_1}\,dt + (2+\sigma_2)^{-1}\int r|u_n|^{2+\sigma_2}\,dt \\ \leqq C_{1,1/4}(\mu^{2+\alpha_1} + \mu^{2+\beta_1})\end{aligned} \tag{3.2}$$

holds for all n.

In particular, we see that (u_n) is bounded in H^1. Then, using Lemma 2.1, the reflexivity of H^1, the Arzela-Ascoli theorem, and a standard diagonal process, it follows that there exists a subsequence of (u_n), still denoted by (u_n), and a $u \in H^1$ such that $u_n \to_w u$ in H^1 and $\sup_{|t| \leqq d} |u(t) - u_n(t)| \to_{n\to\infty} 0$ holds for all $d \geqq 0$. Moreover, by (3.2), Fatou's lemma and the uniform boundedness principle it follows that $u \in S_\mu$. Using the fact that $q_1 \in L^\infty$ and that $q_1(t) \to 0$ as $|t| \to \infty$, we verify that

$$\int q_1|u_n|^{2+\sigma_1}\,dt \to \int q_1|u|^{2+\sigma_1}\,dt.$$

Since $\|u|^{2+\sigma_1} - |u_n|^{2+\sigma_1}| \leqq c|u-u_n|(|u|^{1+\sigma_1} + |u_n|^{1+\sigma_1})$ holds for all n and a constant c, we obtain by Hölder's inequality and Lemma 3.1:

$$\begin{aligned}&\int q_2\|u|^{2+\sigma_1} - |u_n|^{2+\sigma_1}|\,dt \\ &\leqq c\left(\int q_2|u-u_n|^{2+\sigma_1}\,dt\right)^{1/(2+\sigma_1)} \\ &\quad\cdot\left(\int q_2(|u|^{1+\sigma_1} + |u_n|^{1+\sigma_1})^{(2+\sigma_1)/(1+\sigma_1)}\,dt\right)^{(1+\sigma_1)/(2+\sigma_1)} \\ &\leqq c'\left(\int q_2|u-u_n|^{2+\sigma_1}\,dt\right)^{1/(2+\sigma_1)}.\end{aligned}$$

The constant p_1 may be chosen as in condition (i.b).

If $p_1 = 1$, we define $p_1' = (2+\sigma_1)p_1' = \infty$, then we obtain for each constant $R > 0$:

$$\begin{aligned}&\int q_2|u-u_n|^{2+\sigma_1}\,dt \\ &\leqq \|q_2\|_{p_1}\|u-u_n\|^{2+\sigma_1}_{L^{(2+\sigma_1)p_1'}(\{t;|t|<R\})} + \|q_2\|_{L^{p_1}(\{t;|t|\geqq R\})}\|u-u_n\|^{2+\sigma_1}_{(2+\sigma_1)p_1'}.\end{aligned}$$

Since $\|q_2\|_{L^{p_1}(\{t;|t|\geqq R\})} \to 0$ as $R \to \infty$, we see that

$$\int q_2|u-u_n|^{2+\sigma_1}\,dt \to 0 \quad (n \to \infty)$$

and

$$\int q_2|u_n|^{2+\sigma_1}\,dt \to \int q_2|u|^{2+\sigma_1}\,dt.$$

The above results show that

$$I(\mu) \leqq \xi(u) \leqq \liminf \xi(u_n) = I(\mu) < 0.$$

Moreover, (3.2) holds true if we replace u_n by u.

Since $\xi(u) < 0$, it follows that $\|u\|_2 > 0$. Finally, the fact that u_n converges pointwise to u shows that u is nonnegative.

Now, we assume that (ii) and (ii.a)-(ii.c) are fulfilled. The sequence $(u_n) \subset S_\mu$ may be chosen as above. Then, according to Lemma 3.2, we obtain

$$\frac{1}{2}\|u_n'\|_2^2+(2+\sigma_1)^{-1}\int |q_-||u_n|^{2+\sigma_1}\,dt+(2(2+\sigma_2))^{-1}\int r|u_n|^{2+\sigma_2}\,dt \leqq C_{2,\varepsilon}(\mu^{\alpha_2}+\mu^{\beta_2})$$

for $\varepsilon = r_0/(2(2+\sigma_2))$ and all n. Here, the constant $r_0 > 0$ is chosen as in (ii.a). Then, proceeding as above, we obtain the assertion. □

LEMMA 3.6. *Suppose that there exists a constant $\mu > 0$ such that $I(\mu) < 0$. Further, we assume that the function u is chosen according to Lemma* 3.5. *Then* (1.1) *holds in the sense of distributions if*

$$\lambda = \|u\|_2^{-2}\left(\|u'\|_2^2 - \int q|u|^{2+\sigma_1}\,dt + \int r|u|^{2+\sigma_2}\,dt\right).$$

Proof. Since $u \in L^\infty$, we see that $q|u|^{\sigma_1}u$ and $r|u|^{\sigma_2}u$ are locally integrable. Using this fact, it is not difficult to show that for each $\varphi \in C_0^\infty$ the function

$$\varepsilon \to \Phi(\varepsilon) = \xi(\|u\|_2\|u+\varepsilon\varphi\|_2^{-1}(u+\varepsilon\varphi))$$

is differentiable at $\varepsilon = 0$. But $d\Phi(\varepsilon)/d\varepsilon\,|_{\varepsilon=0} = 0$ then implies the assertion. □

LEMMA 3.7. *The constants μ and λ and the function u may be chosen as in Lemma* 3.6. *Then u is positive and continuously differentiable. Moreover, if q and r are continuous, then u is twice continuously differentiable and* (1.1) *holds in the classical sense.*

Proof. The function F may be defined by

$$F(t) = -q(t)u^{1+\sigma_1}(t) + r(t)u^{1+\sigma_2}(t) - \lambda u(t).$$

Since u is bounded, we see that F is locally integrable. Then, for a fixed x, we define

$$U(t) = \int_x^t \int_x^s F(y)\,dy\,ds.$$

Hence, U is continuously differentiable.

For $\varepsilon > 0$, we define $F_\varepsilon = F^*\rho_\varepsilon$, where ρ_ε is a mollifier (cf. [1, p. 29]), and

$$U^\varepsilon(t) = \int_x^t \int_x^s F_\varepsilon(y)\,dy\,ds.$$

Then, it is not difficult to verify that

$$(3.3) \qquad F_\varepsilon \to F \quad \text{and} \quad U^\varepsilon \to U \quad \text{in} \quad L^1_{\text{loc}} \quad \text{as } \varepsilon \to 0.$$

Since $(U^\varepsilon)'' = F_\varepsilon$, we conclude from (3.3) that $U'' = F$ holds in the sense of distributions. Thus, we see that

$$(3.4) \qquad u(t) = U(t) + c_1 t + c_2$$

holds for all t and some constants c_1 and c_2 (cf. [1, Cor. 3.27]). But (3.4) shows that u is continuously differentiable.

Now, suppose that $u(x) = 0$ holds for some x. Since $u \geqq 0$, it follows that $u'(x) = 0$. The function U may be defined as above. Then, we also have $U(x) = 0$ and $U'(x) = 0$. Thus, the constants c_1 and c_2 in (3.4) are zero, and $u(t) = U(t)$ holds for all t.

The function F can be written as $F = Gu$, where G again is locally integrable. Then, for $R > 0$ and all $t \in I_R(x) := (x-R, x+R)$, we obtain

$$u(t) \leqq \int_x^t \int_x^s |G(y)|u(y)\,dy\,ds \leqq R\|G\|_{L^1(I_R(x))} \cdot \sup_{y \in I_R(x)} u(y).$$

Hence, we see that

$$\sup_{y\in I_R(x)} u(y) \leqq R\|G\|_{L^1(I_R(x))} \cdot \sup_{y\in I_R(x)} u(y). \tag{3.5}$$

But (3.5) shows that there is an $R>0$ such that $\sup_{y\in I_R(x)} u(y)=0$. Since u is continuous, the above argument shows that if $u(x)=0$ holds for some x, then $u(x)=0$ holds for all x. However, $u\equiv 0$ contradicts $\|u\|_2>0$.

Finally, we assume that the functions q and r are continuous. Then, we conclude from Lemma 2.1 that F also is continuous and that U is twice continuously differentiable. Thus, (3.4) shows that u is twice continuously differentiable. □

LEMMA 3.8. *The constant λ in Lemma 3.6 always satisfies $\lambda\leqq 0$.*

Proof. For all $t\in(0,1]$ we have $\xi(u)\leqq\xi(tu)$. Hence

$$\lambda=\|u\|_2^{-2}\, d\xi\frac{(tu)}{dt}\bigg|_{t=1}\leqq 0. \qquad □$$

LEMMA 3.9. *Suppose that the constant λ in Lemma 3.6 is negative. Then we have $\|u\|_2=\mu$.*

Proof. If $\|u\|_2<\mu$, then it follows that

$$\lambda=\|u\|_2^{-2}\, d\xi\frac{(tu)}{dt}\bigg|_{t=1}=0. \qquad □$$

LEMMA 3.10. *Suppose that λ is negative and that there exist positive constants t_1 and C such that*

$$q(t)\leqq C \quad \textit{holds for almost all } |t|\geqq t_1.$$

Then, for each $\alpha\in(0,|\lambda|^{1/2})$, there exists a constant C_α such that

$$u(t)\leqq C_\alpha \exp(-\alpha|t|) \textit{ holds for all } t.$$

Proof. From Lemma 3.6 we conclude that

$$\int u'\varphi'\,dt\leqq -|\lambda|\int u\varphi\,dt + C\int u^{1+\sigma_1}\varphi\,dt$$

holds for all nonnegative functions $\varphi\in C_0^\infty(\{t;|t|\geqq t_1\})$.

Now let $\varepsilon\in(0,|\lambda|-\alpha^2)$. Then, since u vanishes at infinity, there exists a $t_2>t_1$ such that $u^{\sigma_1}(t)\leqq\varepsilon/C$ holds for all $|t|\geqq t_2$. Hence, we obtain

$$\int u'v'\,dt\leqq -|\lambda|\int uv\,dt+\varepsilon\int uv\,dt \tag{3.6}$$

for all nonnegative functions $v\in H_0^1(\{t;|t|\geqq t_2\})$.

Since u is bounded, we can find a constant $C_\alpha>0$ such that

$$u(t)\leqq C_\alpha\exp(-\alpha|t|)=:\psi_\alpha(t) \quad \text{holds for all } |t|\leqq t_2+1. \tag{3.7}$$

Because $\psi_\alpha''(t)=\alpha^2\psi_\alpha(t)(t\neq 0)$, it follows from (3.6) that

$$\int (u'-\psi_\alpha')v'\,dt\leqq -\alpha^2\int (u-\psi_\alpha)v\,dt. \tag{3.8}$$

From (3.7), we conclude that $(u-\psi_\alpha)_+\in H_0^1(\{t;|t|\geqq t_2\})$. Inserting $v=(u-\psi_\alpha)_+$ in (3.8) finishes the proof. □

Lemmas 3.9 and 3.10 motivate the question if the constant λ is negative. The first result in this direction is the following.

LEMMA 3.11. *The constants μ and λ and the function u may be chosen as in Lemma 3.6. Moreover, we assume that $\sigma_2 \leqq \sigma_1$ or $r\equiv 0$. Then it follows that $\lambda<0$.*

Proof. Suppose that $\sigma_2\leqq\sigma_1$. Then $\xi(u)<0$ implies that

$$(2+\sigma_1)^{-1}\lambda<\|u\|_2^{-2}\xi(u)<0.$$

In case that $r\equiv 0$, we may assume without restriction that $\sigma_2\leqq\sigma_1$. □

LEMMA 3.12. *If $\sigma_1\geqq 2(2-a)$, then λ is negative.*

Proof. In the following, we take up an idea that we found in [2] (see Lemma 13). Assume that $\lambda\geqq 0$. Then, we obtain

$$\int u'\varphi'\,dt+\mathcal{K}\int t^b u^{1+\sigma_2}\varphi\,dt\geqq 0 \tag{3.9}$$

for all nonnegative functions $\varphi\in C_0^\infty((t_0,\infty))$, where t_0 is defined as in condition (i.c), respectively, (ii.c).

Moreover, via regularization it follows that (3.9) holds for all nonnegative functions $\varphi\in H_0^1((t_0,\infty))$ with compact support.

For $t\geqq t_0$, we define $\psi(t)=Ct^{-1/2}$, where the constant $C\in(0,(3/4\mathcal{K})^{1/\sigma_2}]$ is chosen such that

$$\psi(t)\leqq u(t)\quad\text{holds on } [t_0,t_0+1]. \tag{3.10}$$

Then, it follows that

$$-\psi''(t)+\mathcal{K}t^b\psi^{1+\sigma_2}(t)=C\{\mathcal{K}C^{\sigma_2}t^{b-1/2-\sigma_2/2}-\tfrac{3}{4}t^{-5/2}\}.$$

From the definition of the constant b and the fact that $\sigma_1\geqq 2(2-a)$ we obtain that $b-\frac12-\sigma_2/2\leqq-\frac52$. Hence,

$$-\psi''(t)+\mathcal{K}t^b\psi^{1+\sigma_2}(t)\leqq 0\quad\text{holds for all } t\geqq t_0. \tag{3.11}$$

Since $\psi',\psi^{1+\sigma_2}\in L^2((t_0,\infty))$, we conclude from (3.11) that

$$\int \psi'\varphi'\,dt+\mathcal{K}\int t^b\psi^{1+\sigma_2}\varphi\,dt\leqq 0 \tag{3.12}$$

holds for all nonnegative functions $\varphi\in H_0^1((t_0,\infty))$ with compact support. The function $\zeta\in C_0^\infty$ may be chosen such that $0\leqq\zeta\leqq 1$, $\zeta\equiv 1$ on the unit ball and $\zeta(t)=0$ holds for $|t|\geqq 2$. Moreover, for $n\in\mathbb{N}$, we define $\zeta_n(t)=\zeta(n^{-1}t)$. Then, according to (3.10), we see that the support of the function $(\psi-u)_+\zeta_n$ is a compact subset of (t_0,∞) and that $(\psi-u)_+\zeta_n\in H_0^1((t_0,\infty))$. Moreover, since

$$\int_{t_0}^\infty((\psi-u)_+')^2\zeta_n\,dt=\int_{t_0}^\infty(\psi-u)'(((\psi-u)_+\zeta_n)'-(\psi-u)_+\zeta_n')\,dx,$$

we obtain from (3.9) and (3.12) that

$$\begin{aligned}&\int_{t_0}^\infty((\psi-u)_+')^2\zeta_n\,dt+\mathcal{K}\int_{t_0}^\infty t^b(\psi^{1+\sigma_2}-u^{1+\sigma_2})(\psi-u)_+\zeta_n\,dt\\ &\leqq-\int_{t_0}^\infty(\psi-u)'(\psi-u)_+\zeta_n'\,dt\\ &\leqq\frac1n\|\zeta'\|_\infty\|(\psi-u)'\|_{L^2((t_0,\infty))}\left(\int_{t_0}^{2n}\psi^2(t)\,dt\right)^{1/2}.\end{aligned}\tag{3.13}$$

Since $\int_{t_0}^{2n} \psi^2(t)\,dt = C^2(\log 2n - \log t_0)$, it follows that the right-hand side of (3.13) converges to zero as $n \to \infty$. Hence, we see that

$$\psi(t) \leqq u(t) \quad \text{holds for all } t \geqq t_0. \tag{3.14}$$

But (3.14) contradicts $u \in L^2$. □

Proof of Theorem 1.1. The assertions of Theorem 1.1 now follow from Lemmas 3.4–3.7, 3.9, 3.10, and 3.12. □

4. The case where $\sigma_1 < 2(2-a)$. From Lemma 3.4 it follows that there is a $\mu_0 > 0$ such that $I(\mu) < 0$ holds for all $\mu \in (0, \mu_0]$. Thus, Lemmas 3.5–3.7 show that for each $\mu \in (0, \mu_0]$ there exist a positive function $u_\mu \in S_\mu$ and a constant λ_μ such that $\xi(u_\mu) = I(\mu)$ and (u_μ, λ_μ) solves (1.1).

Since $I(\cdot)$ is a monotone decreasing function on $[0, \mu_0]$, we can find a measurable subset $\mathcal{M}$ of $(0, \mu_0)$ such that $(0, \mu_0) \backslash \mathcal{M}$ has measure zero and $I(\cdot)$ is differentiable on $\mathcal{M}$ (see [6, Thm. 17.12]).

LEMMA 4.1. *For each $\mu \in \mathcal{M}$ we have $I'(\mu) = \mu^{-1}\|u_\mu\|_2^2\lambda_\mu$.*

Proof. Let $\mu \in \mathcal{M}$ and $\varphi(t) = \xi(tu_\mu)$. Then φ is a continuously differentiable function on $(0, \infty)$ satisfying $\varphi(1) = I(\mu)$ and $\varphi'(1) = \lambda_\mu\|u_\mu\|_2^2$. For $\varepsilon > 0$ we can find a $\delta > 0$ such that

$$|\varphi'(t) - \varphi'(1)| \leqq \varepsilon\|u_\mu\|_2^2 \quad \text{holds for } t \in [1-\delta, 1+\delta]. \tag{4.1}$$

From (4.1), we conclude that

$$\varphi'(t) \leqq \varphi'(1) + \varepsilon\|u_\mu\|_2^2 \quad \text{holds for all } t \in (1, 1+\delta]. \tag{4.2}$$

Hence

$$\varphi(t) = \varphi(1) + \int_1^t \varphi'(s)\,ds \leqq I(\mu) + (t-1)(\lambda_\mu + \varepsilon)\|u_\mu\|_2^2.$$

Since $I(t\mu) \leqq \varphi(t)$, we see that

$$\frac{I(t\mu) - I(\mu)}{(t\mu - \mu)} \leqq \mu^{-1}(\lambda_\mu + \varepsilon)\|u_\mu\|_2^2. \tag{4.3}$$

But (4.3) implies $I'(\mu) \leqq \mu^{-1}\|u_\mu\|_2^2\lambda_\mu$.

Next, we conclude from (4.1) that

$$\varphi'(t) \geqq \varphi'(1) - \varepsilon\|u_\mu\|_2^2 \quad \text{holds for all } t \in [1-\delta, 1). \tag{4.4}$$

Hence, we obtain

$$I(t\mu) \leqq \varphi(t) = \varphi(1) - \int_t^1 \varphi'(s)\,ds \leqq I(\mu) + (t-1)(\lambda_\mu - \varepsilon)\|u_\mu\|_2^2$$

and

$$(I(t\mu) - I(\mu))/(t\mu - \mu) \geqq \mu^{-1}(\lambda_\mu - \varepsilon)\|u_\mu\|_2^2 \quad \text{for } t \in [1-\delta, 1). \tag{4.5}$$

Now, (4.5) implies $I'(\mu) \geqq \mu^{-1}\|u_\mu\|_2^2\lambda_\mu$. □

LEMMA 4.2. *There exists a constant C such that*

$$|I(\mu_1) - I(\mu_2)| \leqq C|\mu_1 - \mu_2| \quad \textit{holds for all } \mu_1, \mu_2 \in [0, \mu_0].$$

Proof. Without restriction we may assume that $\mu_1 < \mu_2$. Then, since $I(\mu_2) \leqq I(\mu_1)$

and $I(\mu_1) \leqq \xi((\mu_1/\mu_2)u_{\mu_2})$, it follows that

$$\begin{aligned}|I(\mu_1)-I(\mu_2)| = I(\mu_1)-I(\mu_2) &\leqq \xi\left(\left(\frac{\mu_1}{\mu_2}\right)u_{\mu_2}\right)-\xi(u_{\mu_2})\\ &= \left(\left(\frac{\mu_1}{\mu_2}\right)^2-1\right)\frac{1}{2}\int |u'_{\mu_2}|^2\,dt\\ &\quad+\left(1-\left(\frac{\mu_1}{\mu_2}\right)^{2+\sigma_1}\right)(2+\sigma_1)^{-1}\int q|u_{\mu_2}|^{2+\sigma_1}\,dt\\ &\quad+\left(\left(\frac{\mu_1}{\mu_2}\right)^{2+\sigma_2}-1\right)(2+\sigma_2)^{-1}\int r|u_{\mu_2}|^{2+\sigma_2}\,dt\\ &\leqq \left(1-\left(\frac{\mu_1}{\mu_2}\right)^{2+\alpha_1}\right)(2+\sigma_1)^{-1}\int q_+|u_{\mu_2}|^{2+\sigma_1}\,dt.\end{aligned}$$

Moreover,

$$1-\left(\frac{\mu_1}{\mu_2}\right)^{2+\sigma_1}=(2+\sigma_1)\int_{\mu_1/\mu_2}^{1} t^{1+\sigma_1}\,dt \leqq (2+\sigma_1)\left(1-\left(\frac{\mu_1}{\mu_2}\right)\right)$$

implies that

$$|I(\mu_1)-I(\mu_2)| \leqq \left(1-\left(\frac{\mu_1}{\mu_2}\right)\right)\int q_+|u_{\mu_2}|^{2+\sigma_1}\,dt.$$

Since $\sigma_1<4$, we conclude from Lemmas 3.1 and 3.5 that

$$\int q_+|u_{\mu_2}|^{2+\sigma_1}\,dx \leqq c(\mu_2^{2+\alpha_1}+\mu_2^{2+\beta_1})$$

holds for some constant c, and consequently that

$$|I(\mu_1)-I(\mu_2)| \leqq c|\mu_1-\mu_2|(\mu_0^{1+\alpha_1}+\mu_0^{1+\beta_1}). \qquad \square$$

LEMMA 4.3. *There exists a sequence* (μ_n) *of pairwise distinct constants* $\mu_n \in (0, \mu_0)$ *such that* $\lim_{n\to\infty}\mu_n=0$ *and* $\lambda_{\mu_n}<0$ *holds for all n.*

Proof. Suppose that $\lambda_\mu \geqq 0$ holds for all $\mu \in (0,\mu_0)$. Then, according to Lemma 3.8, $\lambda_\mu=0$ holds for all $\mu\in(0,\mu_0)$.

Now, Lemma 4.1 implies $I'(\mu)=0$ for all $\mu \in \mathcal{M}$. Moreover, Lemma 4.2 shows that $I(\cdot)$ is absolutely continuous on $[0,\mu_0]$.

Hence, from Theorem 18.15 in [6], we conclude that $I(\cdot)$ is constant. But $I(\mu)=I(\mu_0)<0$ for all $\mu\in[0,\mu_0]$ contradicts $I(0)=0$. Thus, there exists a constant $\mu_1\in(0,\mu_0)$ such that $\lambda_{\mu_1}<0$. Next, repeating this procedure, we can find a $\mu_2\in(0,\min(\mu_1,\frac{1}{2}))$ so that $\lambda_{\mu_2}<0$. Moreover, by induction we can show that for each n there is a constant $\mu_n\in(0,\min(\mu_{n-1},1/n))$ so that $\lambda_{\mu_n}<0$. $\square$

LEMMA 4.4. *The sequence* (μ_n) *may be chosen as in Lemma* 4.3. *Then, we have* $\lim_{n\to\infty}\lambda_{\mu_n}=0$.

Proof. Since $\lambda_{\mu_n}<0$, it follows that $\|u_{\mu_n}\|_2=\mu_n$ (see Lemma 3.9) and that $|\lambda_{\mu_n}|\leqq \mu_n^{-2}\int q_+|u_{\mu_n}|^{2+\sigma_1}\,dt$ (see Lemma 3.6). Now, the assertion follows from Lemmas 3.1 and 3.5 and the fact that $\mu_n\to 0$ as $n\to\infty$. $\square$

Proof of Theorem 1.2. Let $u_n=u_{\mu_n}$ and $\lambda_n=\lambda_{\mu_n}$. Then, from Lemmas 3.1 and 3.5, we obtain that

$$\|u_n\|_{H^1}^2+\int |q||u_n|^{2+\sigma_1}\,dt+\int r|u_n|^{2+\sigma_2}\,dt\to 0 \tag{4.6}$$

as $n \to \infty$. Moreover, Lemma 2.1 and (4.6) imply that $\|u_n\|_p \to 0$ if $p \in [2, \infty]$. The remaining assertions follow from the above results. □

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] R. Benguria, H. Brezis, and E. H. Lieb, *The Thomas-Fermi-von Weizsäcker theory of atoms and molecules*, Comm. Math. Phys., 79 (1981), pp. 167-180.

[3] H. Berestycki and P.-L. Lions, *Nonlinear scalar field equations*, I *Existence of a ground state*, Arch. Rational Mech. Anal., 82 (1983), pp. 313-345.

[4] M. S. Berger, *On the existence and structure of stationary states for a nonlinear Klein-Gordon equation*, J. Funct. Anal., 9 (1972), pp. 249-261.

[5] D.-M. Cao, *Positive solution and bifurcation from the essential spectrum of a semilinear elliptic equation on* $\mathbb{R}^N$, Nonlinear Anal. TMA, 15 (1990), pp. 1045-1052.

[6] E. Hewitt and K. Stromberg, *Real and Abstract Analysis*, Springer-Verlag, Berlin, New York, 1975.

[7] W. Rother, *Bifurcation of nonlinear elliptic equations on* $\mathbb{R}^N$, Bull. London Math. Soc., 21 (1989), pp. 567-572.

[8] ———, *Bifurcation of nonlinear elliptic equations on* $\mathbb{R}^N$ *with radially symmetric coefficients*, Manuscripta Math., 65 (1989), pp. 413-426.

[9] ———, *Existence theorems for a nonlinear elliptic eigenvalue problem on* $\mathbb{R}^N$, Nonlinear Anal. TMA, 15 (1990), pp. 381-386.

[10] ———, *The existence of infinitely many solutions all bifurcating from* $\lambda = 0$, Proc. Roy. Soc. Edinburgh, 118 A (1991), pp. 295-303.

[11] ———, *Existence and bifurcation results for a class of nonlinear boundary value problems in* $(0, \infty)$, Comment. Math. Univ. Carolin., 32 (1991), pp. 297-305.

[12] ———, *Nonlinear scalar field equations*, Differential Integral Equations, 5 (1992), pp. 777-792.

[13] H.-J. Rupen, *The existence of infinitely many bifurcation branches*, Proc. Roy. Soc. Edinburgh, 101 A (1985), pp. 307-320.

[14] W. Strauss, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys., 55 (1977), pp. 149-162.

[15] C. A. Stuart, *Bifurcation pour des problèmes de Dirichlet et de Neumann sans valeurs propres*, C. R. Acad. Sci. Paris Sér. A, 288 (1979), pp. 761-764.

[16] ———, *Bifurcation from the continuous spectrum in the* L^2*-theory of elliptic equations on* $\mathbb{R}^N$, in Recent Methods in Nonlinear Analysis and Applications, Proc. of SAFA IV, Liguori, Napoli, 1981, pp. 231-300.

[17] ———, *Bifurcation for Dirichlet problems without eigenvalues*, Proc. London Math. Soc., 45 (1982), pp. 169-192.

[18] ———, *Bifurcation from the essential spectrum*, Lecture Notes in Math. 1017, Springer-Verlag, New York, (1983), pp. 575-596.

[19] ———, *Bifurcation in* $L^p(\mathbb{R})$ *for a semilinear equation*, J. Differential Equations, 64 (1986), pp. 294-316.

[20] ———, *Bifurcation in* $L^p(\mathbb{R}^N)$ *for a semilinear elliptic equation*, Proc. London Math. Soc., 57 (1988), pp. 511-541.

[21] ———, *Bifurcation from the essential spectrum for some non-compact non-linearities*, Math. Meth. Appl. Sci., 11 (1989), pp. 525-542.

[22] J. F. Toland, *Positive solutions of nonlinear elliptic equations—existence and nonexistence of solutions with radial symmetry in* $L^p(\mathbb{R}^N)$, Trans. Amer. Math. Soc., 282 (1984), pp. 335-354.

[23] H.-S. Zhou and X. P. Zhu, *Bifurcation from the essential spectrum of superlinear elliptic equations*, Appl. Anal., 28 (1988), pp. 51-61.

SIAM J. MATH. ANAL.
Vol. 24, No. 3, pp. 681-702, May 1993

009

SOME ERGODIC THEOREMS FOR SOLUTIONS OF HOMOGENEOUS DIFFERENTIAL EQUATIONS*

KRZYSZTOF WYSOCKI†

Abstract. Consider $x'=f(x)$, where f is defined on a cone K with nonempty interior in a Banach space X. This paper studies the conditions on f that ensure that the directions of the solutions, $x(t)/\|x(t)\|$, converge to the same point $u \in K$ no matter what initial conditions are. Problems of this type often arise in some models from population biology. The results are obtained for maps that are homogeneous of degree 1. In the proofs Hilbert's projective metric is used.

Key words. cones, eigenvectors, Hilbert's projective metric, irreducible operators, order-preserving flows, homogeneous differential equations

AMS subject classifications. 34E, 34H, 47B, 47H

Introduction. In this paper we investigate the asymptotic behavior of directions of solutions of homogeneous differential equations.

More precisely, if K is a cone with nonempty interior in a Banach space X, $f: K \to X$ is positively homogeneous of degree 1 and $x(t, \xi)$ denotes a solution of $x'=f(x)$, $x(0)=\xi$, we ask when

$$\frac{x(t, \xi)}{\|x(t, \xi)\|} \to u \in K$$

for all ξ in the interior of K.

This question arises in many places. It appears in some problems in mathematical economics and in some models from population biology; see [9], [14]. In the population biology literature results concerning this type of the behavior of the solutions are called "ergodic theorems." Here the term "ergodic" refers to the behavior of solutions which is independent of the initial conditions.

We will study this question under an additional assumption that the flow generated by $x'=f(x)$ is order-preserving. In recent years several authors have analyzed convergence of bounded trajectories of certain types of differential equations by exploiting order-preserving properties of their flows; see, e.g., Hirsch [4]. In this article we use a simple fact that order-preserving flows which are also homogeneous of degree 1 are nonexpansive with respect to so-called Hilbert's projective metric d.

Our problem is also related to the question about eigenvectors of cone maps which are homogeneous of degree 1. If $f: K \to X$ is homogeneous of degree 1 and $x(t, \xi)/\|x(t, \xi)\| \to u$ for all $\xi \in \operatorname{int} K$, then u is necessarily an eigenvector of f, and if $u \in \operatorname{int} K$, then f has a unique eigenvector u of norm 1 in int K. Observe that if X is finite-dimensional and f is homogeneous of degree 1 with $f(K) \subseteq K$, then an application of Brouwer's fixed point theorem implies that f has a normalized eigenvector in K. We will consider two possibilities: convergence of directions to an eigenvector in int K and to an eigenvector in ∂K.

In § 1 we recall a few facts about cones and linear cone maps. In particular, we define Hilbert's projective metric d and list some of their properties.

The fact that d is a metric on $S_0=\{x \in \operatorname{int} K, \|x\|=1\}$ and that topologies of S_0 induced by d and $\|\cdot\|$ are identical makes Hilbert's projective metric a useful tool in our considerations.

* Received by the editors December 26, 1990; accepted for publication (in revised form) August 31, 1992.

† School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

In § 2 we give proofs of main results. Convergence of directions to an eigenvector in int K is considered in Theorem 2.1. The following is a special case of this theorem: Suppose that $f: K \to \mathbf{R}^n$ is homogeneous of degree 1 and the flow of $x' = f(x)$ is order-preserving. Assume that f has an eigenvector u of norm 1 in int K and that $f'(u)$ exists, and $f'(u) + \alpha I$ is a nonnegative irreducible matrix for some $\alpha > 0$. Then for any $x \in \operatorname{int} K$, $x(t, x_0)$ is defined in int K for all $t \geqq 0$ and $x(t, x_0)/\|x(t, x_0)\| \to u$.

The second possibility, convergence to an eigenvector in ∂K, is considered in Theorem 2.1. This case presents some difficulties. A complication here is that f is defined only on K and may not have a Fréchet derivative at any point of ∂K. Fréchet differentiability of f at an eigenvector u is replaced by the existence of a Gateaux derivative at u. In fact, many maps of interest have a Gateaux derivative at points of ∂K.

In § 3 we illustrate an application of these results to particular classes of maps which were introduced and analyzed by Nussbaum in [7] and [8]. Finally, let us mention some related problems which are not considered here. We can consider the time-dependent differential equation $x' = f(t, x)$, where $f: \mathbf{R}_+ \times K \to X$. In this situation there are no obvious candidates for asymptotic behavior of solutions, but we can ask for a condition on f which guarantees that any two solutions behave in a similar way. For instance, when

$$\left\| \frac{x(t, x_0)}{\|x(t, x_0)\|} - \frac{x(t, x_1)}{\|x(t, x_1)\|} \right\| \to 0 \quad \text{for all } x_0, x_1 \in \operatorname{int} K$$

or when any two solutions are asymptotically proportional, i.e.,

$$\frac{x_i(t, x_0)}{x_i(t, x_1)} \to c_i,$$

where c_i is a constant independent of initial condition and $x_i(t, \cdot)$ is the ith component of $x(t, \cdot)$ if $X = \mathbf{R}^n$? These questions were studied by Birkhoff and Kotlin in [1] for systems of linear differential equations and asymptotic proportionality by Thieme in [13] for sublinear difference and differential equations.

In [11], Schanbacher obtained results about behavior

$$\frac{T(t)f}{\|T(t)f\|} \quad \text{as } t \to \infty,$$

where $\{T(t)\}$ is a strongly continuous semigroup of positive linear operators on a Banach lattice E and f an element of a positive cone in E.

1. Preliminaries. In this section we state some results on cones in Banach spaces and so-called Hilbert projective metric.

Let X be a real Banach space. A subset K of X is a cone if K is a closed, convex set such that $\lambda K \subseteq K$, for all $\lambda \geqq 0$, and $K \cap (-K) = \{0\}$. A cone K induces a partial ordering on X by $x \leqq y$ if and only if $y - x \in K$. We shall write $x < y$ if $x \leqq y$ but $x \neq y$, while $x \ll y$ indicates that $y - x \in \operatorname{int} K$. Two elements $x, y \in K$ are comparable if there exist positive numbers α, β such that

$$\alpha x \leqq y \leqq \beta x.$$

A cone $K \subseteq X$ is called normal if there is a constant $M > 0$ such that for any $x, y \in K$ if $x \leqq y$ then $\|x\| \leqq M\|y\|$.

By K^* we will denote dual cone, i.e.,

$$K^* = \{\psi \in X^*;\ \psi(x) \geqq 0 \text{ for all } x \in K\}.$$

It is easy to see that if $x \in \operatorname{int} K$, then $\psi(x) > 0$ for all $\psi \in K^*$ and that $x \in \partial K$ if and only if $\psi(x) = 0$ for some $\psi \in K^*$.

Throughout this paper we will always assume that the interior of K, $\operatorname{int} K$, is nonempty and that K is normal.

If x and y are comparable, following notation in [2], we define numbers $m(y/x)$ and $M(y/x)$ by

$$m\left(\frac{y}{x}\right) = \sup\{\alpha > 0;\ \alpha x \leqq y\},$$

$$M\left(\frac{y}{x}\right) = \inf\{\beta > 0;\ y \leqq \beta x\}.$$

Definition 1.1. If x and y are comparable elements of K and $\alpha = m(y/x)$ and $\beta = M(y/x)$, the Hilbert projective metric $d(x, y)$ is defined by

$$d(x, y) := \log \frac{\beta}{\alpha}.$$

If x and y are nonzero elements of K which are noncomparable, we set $d(x, y) = \infty$.

It is not difficult to show, see, e.g., [2], that for all x, y, z in $K \backslash \{0\}$ we have that

$$\begin{aligned} &d(x, z) \leqq d(x, y) + d(y, z),\\ &d(x, y) = d(y, x),\\ &d(x, y) = 0 \Leftrightarrow y = \alpha x \quad \text{for some } \alpha > 0, \quad \text{and}\\ &d(\alpha x, \beta y) = d(x, y) \quad \text{for all } \alpha, \beta > 0. \end{aligned}$$

It follows that d is a pseudometric on $\operatorname{int} K$ and a metric when restricted to $S_0 := \{v \in \operatorname{int} K;\ \|v\| = 1\}$. Moreover, there is a constant $M > 0$ such that

$$\|x - y\| \leqq M[\exp d(x, y) - 1]$$

for all $x, y \in S_0$, and if $u \in \operatorname{int} K$ and $B_r(u) \in \operatorname{int} K$, then

$$d(x, u) \leqq \log\left(\frac{r + \|x - u\|}{r - \|x - u\|}\right) \quad \text{for all } x \in B_r(u).$$

The above inequalities imply that $\|\cdot\|$ and d induce the same topology on S_0. In [14], Thompson has introduced an interesting variant of Hilbert's projective metric. If $x, y \in K \backslash \{0\}$ are comparable, Thomson defines $\bar{d}(x, y)$ by

$$\bar{d}(x, y) = \max\left\{\log M\left(\frac{y}{x}\right), \log M\left(\frac{x}{y}\right)\right\},$$

and if $x, y \in K \backslash 0$ are not comparable, $\bar{d}(x, y) = \infty$.

The restriction of $\bar{d}$ to $\operatorname{int} K$ is a metric and $\|\cdot\|$ and $\bar{d}$ give the same topology on $\operatorname{int} K$. For more information about Hilbert's projective metric and its applications we send the interested reader to Bushell [2], [3] and Nussbaum [7], [8].

We will also need few results concerning linear cone maps.

If $L: X \to X$ is a bounded linear map, L has a natural extension $\tilde{L}$ to the complexification $\tilde{X}$ of X, i.e.,

$$\begin{aligned} &\tilde{X} = \{x + iy;\ x, y \in X\},\\ &\|x + iy\| := \sup_{\theta \in [0, 2\pi]} \|x \cos \theta + y \sin \theta\|, \quad \text{and}\\ &\tilde{L}(x + iy) = Lx + iLy. \end{aligned}$$

We will assume that whenever statements about spectra of L are made they refer to $\tilde{L}$ defined on $\tilde{X}$.

By $r(L)$ we will denote the spectral radius of L and by $r_e(L)$ the essential spectral radius of L. The essential spectral radius of L, $r_e(L)$, can be defined (see [6]) by

$$r_e(L)=\lim_{n\to\infty}(q(L^n))^{1/n},$$

where

$$q(L)=\inf\{\|L+K\|;\ K \text{ is a compact linear map}\}.$$

It is proved in [6] that if $L\in L(X)$ and $L(K)\subseteq K$, and $r_e(L)<r(L)$, then L has an eigenvector $u\in K$ with an eigenvalue $r=r(L)$ and L^* has an eigenvector $u^*\in K^*$ with the same eigenvalue $r=r(L)$.

If L is compact this result is the Krein-Rutman theorem.

Moreover, if $\lambda\in\sigma(L)$ and $|\lambda|>r_e(L)$, then λ is an eigenvalue of finite algebraic multiplicity and λ is an isolated point of $\sigma(L)$.

DEFINITION 1.2. Let $L\in L(X)$ be such that $L(K)\subseteq K$. The map L is called irreducible if $(\lambda I-L)^{-1}(x)\in \operatorname{int} K$ for any $x\in K\backslash\{0\}$ and any $\lambda>r(L)$.

Remark 1.1. One can define an irreducible map even if int K is empty (see [10]). The definition in [10] agrees with the one above when int K is nonempty.

If K is the standard cone in $\mathbf{R}^n$, i.e., $K=\{x;\ x_i\geqq 0 \text{ for } 1\leqq i\leqq n\}$ and L is a nonnegative matrix the above definition is equivalent to the following one: L is irreducible if for each pair (i,j) with $1\leqq i,j\leqq n$ there is an integer p such that the entry in row i and column j of L^p is positive.

The next result provides information about uniqueness of positive eigenvectors of L and L^*. In the special case, $X=R^n$, it is a classical result of Perron and Frobenius.

THEOREM 1.1 [10]. *Let $L\in L(X)$ be irreducible, and $r=r(L)>0$ is a pole of $(zI-L)^{-1}$.*

Then we have

(1) *r is a pole of order* 1; *and*

(2) *L and L^* have eigenvectors $u\in \operatorname{int} K$ and $u^*\in \operatorname{int} K^*$, respectively, with eigenvalue r;*

(3) $\dim\ker(rI-L)=1$.

2. Some ergodic theorems. Let us consider the differential equation

$$\begin{aligned} x'(t)&=f(x),\\ x(0)&=x_0\in \operatorname{int} K. \end{aligned} \tag{1}$$

We are interested in the conditions on f which will guarantee that $x(t,x_0)/\|x(t,x_0)\|$ converges to the same point of int K, no matter what x_0 is. We start by proving a simple lemma which provides the candidate for a limit of $x(t,x_0)/\|x(t,x_0)\|$.

LEMMA 2.1. *Assume that $f\colon \operatorname{int} K\to X$ is a locally Lipschitz map which can be continuously extended to K. Assume that f is positively homogeneous of degree* 1, *i.e., $f(tx)=tf(x)$ for all $t>0$ and $x\in K$, and that any solution of* (1) *is defined for all $t\geqq 0$ and stays in* int K. *If*

$$\frac{x(t,x_0)}{\|x(t,x_0)\|}\to u,\quad t\to\infty\quad \textit{for all } x_0\in \operatorname{int} K,$$

then u is an eigenvector of f.

Proof. Observe that by homogeneity of f and uniqueness of solutions, we have

$$x(t,\lambda x_0)=\lambda x(t,x_0) \tag{2}$$

for any $x_0 \in \operatorname{int} K$ and $\lambda > 0$. Take $\psi \in K^*\backslash\{0\}$ so that $\psi(u) = 1$. Then it follows that

$$\frac{x(t, x_0)}{\psi(x(t, x_0))} \to u. \tag{3}$$

From continuous dependence of solutions on initial conditions and (2) and (3) we get

$$\frac{x(t, u)}{\psi(x(t, u))} = \lim_{s\to\infty} \frac{\left[x\left(t, \dfrac{x(s, x_0)}{\psi(x(s, x_0))}\right)\right]}{\left[\psi\left(x\left(t, \dfrac{x(s, x_0)}{\psi(x(s, x_0))}\right)\right)\right]}$$

$$= \lim_{s\to\infty} \frac{x(t+s, x_0)}{\psi(x(t+s, x_0))} = u \quad \text{for any } t \geqq 0.$$

Hence $x(t, u) = \psi(x(t, u))u$ for any $t \geqq 0$.

After differentiating the above equality we obtain

$$f(x(t, u)) = x'(t, u) = \psi(x'(t, u))u$$
$$= \psi(f(x(t, u)))u.$$

Now our claim follows by taking the limit as $t \to 0^+$. □

The following definition will be useful in our considerations.

Definition 2.1. Let X be a Banach space with ordering induced by a cone K and $f: K \to X$ be locally Lipschitz. We say that the flow generated by (1) is order-preserving if $x(t, x_0) \leqq x(t, x_1)$ for all $t \geqq 0$, whenever $x_0 \leqq x_1$.

Remark 2.1. In practice it can be difficult to check that the flow is order-preserving. When $f: K \to X$ satisfy: whenever $x \leqq y$ with $y - x \in \partial K$ there is $\psi \in K^*\backslash\{0\}$ for which $\psi(y-x) = 0$ and $\psi(f(x)) \leqq \psi(f(y))$ then, as it is shown in [5], the flow of (1) is order-preserving.

If K is the standard cone in $\mathbf{R}^n$ and $f: K \to \mathbf{R}^n$, then the above condition implies that the flow of (1) preserves ordering if whenever x, y are in K with $x_j \leqq y_j$ for all $1 \leqq j \leqq n$ and $x_i = y_i$ for some $1 \leqq i \leqq n$, then $f_i(x) \leqq f_i(y)$, where f_i denotes the ith component of f.

Our interest of order-preserving flows stems from the following simple lemma.

Lemma 2.2. *Suppose that $f: \operatorname{int} K \to X$ is locally Lipschitz and positively homogeneous of degree* 1. *If the flow generated by* (1) *is order-preserving, then*

$$d(x(t, x_0), x(t, x_1)) \leqq d(x_0, x_1) \quad \text{and} \quad \bar{d}(x(t, x_0), x(t, x_1)) \leqq \bar{d}(x_0, x_1)$$

for all $x_0, x_1 \in \operatorname{int} K$ and $t \geqq 0$.

Proof. Let $x_0, x_1 \in \operatorname{int} K$ and $\alpha = m(x_1/x_0)$ and $\beta = M(x_1/x_0)$. Then, by definition of Hilbert's projective metric, we have that

$$\alpha x_0 \leqq x_1 \leqq \beta x_0 \quad \text{and} \quad d(x_0, x_1) = \log\frac{\beta}{\alpha}.$$

Since the flow is order-preserving and f is homogeneous of degree 1,

$$\alpha x(t, x_0) \leqq x(t, x_1) \leqq \beta x(t, x_0).$$

This implies that

$$\alpha \leqq m\left(\frac{x(t, x_1)}{x(t, x_0)}\right) \leqq M\left(\frac{x(t, x_1)}{x(t, x_0)}\right) \leqq \beta,$$

and so

$$d(x(t, x_0), x(t, x_1)) \leqq d(x_0, x_1).$$

The proof for $\bar{d}$ is similar. □

Using this lemma we prove the following.

LEMMA 2.3. *Let $f\colon K \to X$ be continuous and $f_{|\text{int } K}$ locally Lipschitz. Assume that f is positively homogeneous of degree 1 and that the flow of (1) is order-preserving. Define $S = \{v \in K;\ \|v\| = 1\}$ and $S_0 = \{v \in \text{int } K;\ \|v\| = 1\}$. Suppose that there is $u \in S$ and $\delta > 0$ such that*

$$\lim_{t\to\infty} \left\| \frac{x(t, x_0)}{\|x(t, x_0)\|} - u \right\| = 0$$

for all $x_0 \in W = \{v \in S_0;\ \|v - u\| < \delta\}$.

Then it follows that

$$\lim_{t\to\infty} \left\| \frac{x(t, x_0)}{\|x(t, x_0)\|} - u \right\| = 0$$

for all $x_0 \in S_0$.

Proof. Let

$$U = \left\{ x_0 \in S_0;\ \lim_{t\to\infty} \left\| \frac{x(t, x_0)}{\|x(t, x_0)\|} - u \right\| = 0 \right\}.$$

Because $U \neq \emptyset$ and S_0 is connected it suffices to show that U is open and closed in the norm topology of S_0. To see that U is open we take $y \in U$ and $t_0 > 0$ such that $x(t_0, y)/\|x(t_0, y)\| \in W$. Let $\varepsilon > 0$ be such that $\{z \in S_0;\ d(x(t_0, y), z) < \varepsilon\} \subset W$. Define $V = \{z \in S_0;\ d(y, z) < \varepsilon\}$. Since the topologies induced by $\|\cdot\|$ and d on S_0 are identical, V is an open neighborhood of y in S_0. It suffices to show that $V \subseteq U$. Take $y_1 \in V$. Then $d(x(t, y), x(t, y_1)) \leqq d(y, y_1) < \varepsilon$ and, by definition of $\varepsilon > 0$, it follows that $a = x(t_0, y_1)/\|x(t_0, y_1)\| \in W$. Hence

$$\lim_{s\to\infty} \left\| \frac{x(s, a)}{\|x(s, a)\|} - u \right\| = 0$$

and, since

$$\frac{x(s, a)}{\|x(s, a)\|} = \frac{x(s + t_0, y_1)}{\|x(s + t_0, y_1)\|}$$

we get that

$$\lim_{t\to\infty} \left\| \frac{x(t, y_1)}{\|x(t, y_1)\|} - u \right\| = 0.$$

This proves $V \subseteq U$.

To see that U is closed, let $\{x_n\} \subseteq U$ be such that $\lim_{n\to\infty} \|x_n - x_0\| = 0$ for some $x_0 \in S_0$. Then $\lim_{n\to\infty} d(x_n, x_0) = 0$. We know that there is a constant $M > 0$ such that

$$\|x - y\| \leqq M[e^{d(x, y)} - 1] \quad \text{for all } x, y \in S_0.$$

Take $\varepsilon > 0$ and $n_0 \in N$ so that

$$e^{d(x_0, x_{n_0})} - 1 \leqq \frac{\varepsilon}{2M}.$$

Since $x_{n_0} \in U$, there is $t_0 > 0$ so that

$$\left\| \frac{x(t, x_{n_0})}{\|x(t, x_{n_0})\|} - u \right\| < \frac{\varepsilon}{2} \quad \text{for all } t \geqq t_0.$$

Then

$$\left\| \frac{x(t, x_0)}{\|x(t, x_0)\|} - u \right\| \leqq \left\| \frac{x(t, x_0)}{\|x(t, x_0)\|} - \frac{x(t, x_{n_0})}{\|x(t, x_{n_0})\|} \right\| + \frac{\varepsilon}{2}$$

$$\leqq M[e^{d(x(t,x_0),x(t,x_{n_0}))} - 1] + \frac{\varepsilon}{2}$$

$$\leqq M[e^{d(x_0,x_{n_0})} - 1] + \frac{\varepsilon}{2} \leqq \varepsilon.$$

Hence $x_0 \in U$ and U is closed. □

Note that if $\Sigma_0 = \{v \in \operatorname{int} K;\ \psi(v) = 1\}$ and $S_0 = \{v \in \operatorname{int} K;\ \|v\| = 1\}$, then the map $S_0 \ni v \to t_v v \in \Sigma_0$, where t_v is a unique $t > 0$ so that $t_v v \in \Sigma_0$, defines an isometry between (Σ_0, d) and (S_0, d). This implies that topologies induced by $\|\cdot\|$ and d on Σ_0 are the same.

Now, if $\Sigma = \{v \in K;\ \psi(v) = 1\}$ and if $u \in \Sigma$ and $\delta > 0$ are such that

$$\lim_{t\to\infty} \left\| \frac{x(t, x_0)}{\psi(x(t, x_0))} - u \right\| = 0 \quad \text{for all } x_0 \in W = \{v \in \Sigma_0;\ \|v - u\| < \delta\},$$

then a minor modification of the proof of Lemma 2.3 shows that

$$\lim_{t\to\infty} \left\| \frac{x(t, x_0)}{\psi(x(t, x_0))} - u \right\| = 0 \quad \text{for all } x_0 \in \Sigma_0.$$

Now we state our result.

THEOREM 2.1. *Let $f: \operatorname{int} K \to X$ be continuous and $f_{|\operatorname{int} K}$ locally Lipschitz and positively homogeneous of degree* 1. *Assume that the flow of* (1) *is order-preserving and that $f(u) = \lambda u$ for some $u \in \operatorname{int} K$ with $\|u\| = 1$ and $\lambda \in \mathbf{R}$. Assume that f is differentiable at u and that there is a constant α such that $A = (\alpha I + f'(u))$ sends K into itself, $r_e(A) < r(A)$, and A is irreducible. Finally, suppose that* $\sup_{v \in S_0} \|f(v)\| < \infty$.

Then for any $x_0 \in \operatorname{int} K$, $x(t, x_0)$ is defined for all $t \geqq 0$ and $x(t, x_0) \in \operatorname{int} K$, and

$$\frac{x(t, x_0)}{\|x(t, x_0)\|} \to u \quad \textit{as } t \to \infty.$$

In the proof of this theorem it will be more convenient to use $x(t, x_0)/\psi(x(t, x_0))$, where $\psi \in K^* \backslash \{0\}$ is such that $\psi(u) = 1$. Observe that, by remarks following the proof of Lemma 2.3, in order to prove Theorem 2.1 it suffices to find $\delta > 0$ such that $x(t, x_0)/\psi(x(t, x_0)) \to u$ for all $x_0 \in \Sigma_0$ with $\|x_0 - u\| < \delta$.

For the proof we shall need a few more lemmas.

LEMMA 2.4. *Let $L \in L(X)$ be such that $L(K) \subseteq K$ and $Lu \leqq ru$ for some $u \in \operatorname{int} K$ and $r > 0$. Then the spectral radius of L, $r(L)$, is less than or equal to r.*

Proof. Define $X_u = \{x \in X;\ -\alpha u \leqq x \leqq \alpha u$ for some $\alpha\}$ and $\|x\|_u = \inf\{\alpha \geqq 0;\ -\alpha u \leqq x \leqq \alpha u\}$. Then X_u is a normed space with the norm $\|\cdot\|_u$, and since K is normal and $u \in \operatorname{int} K$, we have that $X_u = X$ and the norms $\|\cdot\|$, $\|\cdot\|_u$ are equivalent. It follows that

$$r(L) = \lim_{n\to\infty} \|L^n\|_u^{1/n}.$$

By definition of $\|\cdot\|_u$, we have

$$-\|x\|_u u \leqq x \leqq \|x\|_u u \quad \text{for all } x \in X.$$

Since $L(K) \subseteq K$ and $Lu \leqq ru$, iteration of the above inequalities gives

$$-\|x\|_u r^n u \leqq L^n x \leqq \|x\|_u r^n u \quad \text{for all } n \in N.$$

Hence $\|L^n\|_u \leqq r^n$ and the assertion follows. □

For the proof of the next lemma see [7].

LEMMA 2.5. *Let $L \in L(X)$ be a bounded linear operator such that $L(K) \subseteq K$ and $r(L) \leqq r$. Assume that $Lu \leqq ru$ for some $u \in K\{0\}$. Let $\psi \in K^*$ be so that $\psi(u) = 1$ and $\Lambda x = Lx - \psi(Lx)u$. Then $r(\Lambda) \leqq r$.*

Recall that if $L \in L(X)$, then the modulus of stability of L is defined by

$$s(L) = \sup\{\operatorname{Re}\lambda;\ \lambda \in \sigma(L)\}.$$

LEMMA 2.6. *Let $L \in L(X)$ be such that $L(K) \subseteq K$ and $Lu \leqq ru$ with $u \in \operatorname{int} K$ and $r > 0$. Assume that $r_e(L) < r(L)$ and that L is irreducible. If $\Lambda x = Lx - \psi(Lx)u$, $x \in X$, where $\psi \in K^*$ is such that $\psi(u) = 1$, then*

$$s(\Lambda) < r.$$

Proof. Lemmas 2.4 and 2.5 show that $r(\Lambda) \leqq r$. Hence it suffices to show that $r(\Lambda) < r$. The proof proceeds by contradiction, so we assume that $r(\Lambda) = r$.

Since $x \mapsto \psi(Lx)u$ is a compact linear map, it follows that $r_e(\Lambda) = r_e(L)$ and

$$r_e(\Lambda) < r(\Lambda) = r.$$

Thus r is an eigenvalue of Λ. Denote by v a corresponding eigenvector. There are two cases to consider, namely:

(1) $Lu \leqq ru$ but $Lu \neq ru$; and

(2) $Lu = ru$.

Case (1). We will show that if $Lu \leqq ru$ but $Lu \neq ru$, then $r(L) < r$. It is not difficult to see that if $Lw = \mu w$ for some $w \in K\backslash\{0\}$ and $\mu > 0$, and $Lu \leqq ru$, then $\mu \leqq r$. Now, take $\lambda > r(L)$ and define $x = ru - Lu \in K\backslash\{0\}$ and $y = (\lambda I - L)^{-1}u$.

Because L is irreducible, $y \in \operatorname{int} K$ and

$$(\lambda I - L)^{-1}x = r(\lambda I - L)^{-1}u - L(\lambda I - L)^{-1}u = ry - Ly \in \operatorname{int} K.$$

Hence there is $\varepsilon > 0$ such that

$$Ly \leqq (r - \varepsilon)y.$$

Because $r_e(L) < r(L)$, L has an eigenvector in K with an eigenvalue $r(L)$, which, in view of the remark above, satisfies $r(L) \leqq r - \varepsilon$.

On the other hand,

$$Lv - \psi(Lv)u = rv,$$

and, since $r(L) < r$, we have $v = -\psi(Lv)(rI - L)^{-1}(u)$.

Let $z = (rI - L)^{-1}u$. Then $\psi(z) > 0$ and $0 = \psi(v) = -\psi(Lv)\psi(z)$. This implies that $Lv = rv$ and $r(L) = r$, a contradiction.

Case (2). We know that L^* has an eigenvector w^* corresponding to r and $w^*(u)=1$.

Let

$$Y:=\{x\in X;\ \psi(x)=0\}\quad\text{and}$$
$$Y_1:=\{x\in X;\ w^*(x)=0\}.$$

Define $\Lambda_1: X\to X$ by

$$\Lambda_1:=Lx-w^*(Lx)u.$$

Since X is a direct sum of $\{tu;\ t\in\mathbf{R}\}$ and of Y and both of these subspaces are invariant under Λ, we have that

$$\sigma(\Lambda)=\sigma(\Lambda\,|\,Y)\cup\{0\}.$$

Let $B: Y\to Y_1$ and $C: Y_1\to Y$ be given by

$$B(y)=y-w^*(y)u\quad\text{and}$$
$$C(z)=z\psi(z)u.$$

Then $CB=I_{|Y}$, $BC=I_{|Y_1}$ and $CLB=\Lambda$. It follows that

$$\sigma(\Lambda\,|\,Y)=\sigma(L\,|\,Y_1). \tag{4}$$

As the algebraic multiplicity of r is equal to 1 and $u\notin Y_1$, we find by (4) that

$$\sigma(\Lambda\,|\,Y)=\{\lambda;\ \lambda\in\sigma(L),\ \lambda\neq r\}.$$

Hence $r\notin\sigma(\Lambda)$ and again we have a contradiction. □

After these preliminary lemmas we are ready to prove Theorem 2.1.

Proof of Theorem 2.1. Observe that without loss of generality we can assume that $f(u)=0$, i.e., $\lambda=0$. In fact, if $g(x)=-\lambda x+f(x)$ and $y(t,x_0)=e^{-\lambda t}x(t,x_0)$, then $g(u)=0$, g has the same properties as f and $y'(t,x_0)=g(y(t,x_0))$.

Homogeneity of f implies that

$$x(t,\alpha x_0)=\alpha x(t,x_0)\quad\text{for any }\alpha>0\quad\text{and}\quad x_0\in\operatorname{int}K. \tag{5}$$

Take $x_0\in\operatorname{int}K$ and let $\beta>0$ be such that $u\leqq\beta x_0$. Because the flow of (1) is order-preserving and f has linear growth, we have that

$$u\leqq\beta x(t,x_0)\quad\text{for } t\geqq 0$$

and so $x(t,x_0)\in\operatorname{int}K$ for $t\geqq 0$.

Let $\psi\in K^*\backslash\{0\}$ be so that $\psi(u)=1$ and $\Sigma_0=\{v\in\operatorname{int}K;\ \psi(v)=1\}$. It is more convenient to work with

$$y(t,x_0)=\frac{x(t,x_0)}{\psi(x(t,x_0))}\quad\text{than}\ \frac{x(t,x_0)}{\|x(t,x_0)\|}.$$

Then $y(t,x_0)$ satisfies

$$\begin{aligned}&y'(t,x_0)=f(y(t,x_0))\psi(f(y(t,x_0)))y(t,x_0)\\&y(0,x_0)=\frac{x_0}{\psi(x_0)}.\end{aligned}\tag{6}$$

Moreover, $y(t,x_0)\in\operatorname{int}K$ for all $t\geqq 0$.

We claim that $y(t,x_0)\to u$ as $t\to\infty$ for all $x_0\in\operatorname{int}K$.

Let $Y=\{z\in X;\ \psi(z)=0\}$. Note that if $x\in X$, then it can be written as

$$x=x-\psi(x)u+\psi(x)u,$$

where $x-\psi(x)u\in Y$. Hence $X=X_0\oplus Y$, where $X_0=\text{span}\,(u)$. Write $y(t,x_0)=z(t)+\alpha(t)u$, with $z(t)\in Y$ and $\alpha(t)\in\mathbf{R}$.

Since $\psi(y(t,x_0))=1$, we see that $\alpha(t)=1$, and so

$$y(t,x_0)=z(t)+u, \qquad t\geqq 0.$$

Of course, $z(t)$ satisfies

$$\tag{7} \begin{aligned} z'(t)&=h(z(t)),\\ z(0)&=\frac{x_0}{\psi(x_0)}-u, \end{aligned}$$

where $h(z)=f(u+z)-(u+z)\psi(f(u+z))$. Thus it is enough to show that $z(t)\to 0$, $t\to\infty$. The equation (7) can be written as

$$\tag{8} \begin{aligned} z'(t)&=Bz(t)+g(z(t)),\\ z(0)&=z_0=\frac{x_0}{\psi(x_0)}-u, \end{aligned}$$

with $g: Y\to Y$ satisfying

$$\frac{\|g(z)\|}{\|z\|}\to 0 \quad \text{as } \|z\|\to 0$$

and

$$Bz=h'(0)=f'(u)z-\psi(f'(u)z):=Lz-\psi(Lz)u.$$

By the variation of constants formula solutions of (8) take the form

$$\tag{9} z(t)=e^{Bt}z_0+\int_0^t e^{B(t-s)}g(z(s))\,ds.$$

Let $\Lambda x=Ax\psi(Ax)u$ with $A=\alpha I+L$. Note that, since $f(u)=0$ and f is homogeneous of degree 1, we have that

$$Lu=f(u)=0.$$

Since $Au=\alpha u$, $u\in\text{int}\,K$, Lemma 2.2 implies that $r(A)\leqq\alpha$ and since $r_e(A)<r(A)$, Lemma 2.4 shows that $s(\Lambda)<\alpha$. Because $\Lambda z=(\alpha I+B)z$ for any $z\in Y$ we see that $\sigma(\Lambda\,|\,Y)=\sigma(B)+\alpha$. This shows that $s(B)=\mu<0$. Hence $\|e^{Bt}\|\leqq Me^{\mu_0 t}$ for some M and $\mu<\mu_0<0$.

From this, by Gronwall's inequality applied to (9), we obtain

$$\|z(t,z_0)\|\leqq M_1\,e^{\mu_1 t},$$

where $M_1>0$ and $\mu_1<0$ for z_0 close to zero.

Hence $z(t,z_0)\to 0$ for $\|z_0\|<\delta_0$, where δ_0 is small. This implies that

$$\frac{x(t,x_0)}{\psi(x(t,x_0))}\to u \quad \text{for any } x_0\in\{v\in\Sigma_0;\ \|v-u\|<\delta\}$$

with δ small. Application of Lemma 2.3 finishes the proof. □

In $\mathbf{R}^n$, Theorem 2.1 takes the following simple form.

COROLLARY 2.1. *Let K be the standard cone in $\mathbf{R}^n$ and f: int $K\to\mathbf{R}^n$ be a locally Lipschitz map. Assume that the flow generated by f is order-preserving, that f is positively* *homogeneous of degree* 1 *and* $f(u)=\lambda u$ *for some* $u\in\operatorname{int}K$ *with* $\|u\|=1$ *and* $\lambda\in\mathbf{R}$. *Finally, assume that f is differentiable at u and if*

$$L:=f'(u)=\left[\frac{\partial f_i}{\partial x_j}(u)\right],$$

then $A:=\alpha I+L$ *is an irreducible matrix for some* $\alpha>0$. *Then*

$$\lim_{t\to\infty}\frac{x(t,x_0)}{\|x(t,x_0)\|}=u$$

for all $x_0\in\operatorname{int}K$.

Proof. Compactness and the Perron-Frobenius theorem imply that $r_e(A)=0$ and $r(A)>0$. The other assumptions of Theorem 2.1 are trivially satisfied. □

Remark 2.2. Observe that Theorem 2.1 and Corollary 2.1, and Lemma 2.3 remain true if we replace the assumption that the flow of (1) is order-preserving by a weaker, namely that the flow is nonexpansive with respect to Hilbert's projective metric d.

If in Theorem 2.1, $f(u)=0$, i.e., $\lambda=0$, then any point of $\{\alpha u;\ \alpha\geqq 0\}$ is equilibrium and we can ask about convergence of $x(t,x_0)$ for any $x_0\in\operatorname{int}K$.

PROPOSITION 2.1. *Assume that f satisfies all assumptions of Theorem* 2.1. *and that* $\lambda=0$.

Then there is a continuous function $\gamma:\operatorname{int}K\to\mathbf{R}_+$ *such that for any* $x_0\in\operatorname{int}K$

$$\|x(t,x_0)-\gamma(x_0)u\|\to 0\quad \textit{as } t\to\infty.$$

Moreover, γ *is order-preserving and homogeneous of degree* 1.

Proof. Recall that the topologies induced by $\|\cdot\|$ and Thompson's metric $\bar d$ on $\operatorname{int}K$ are identical. Therefore, it suffices to prove the result for the metric $\bar d$. Let $\{t_n\}$ be any increasing sequence converging to infinity.

It follows from Theorem 2.1 that

$$\lim_{n\to\infty}\frac{x(t_n,x_0)}{\psi(x(t_n,x_0))}=u \tag{10}$$

for any $x_0\in K$, where $\psi\in K^*\backslash\{0\}$ is such that $\psi(u)=1$.

Since $x(t,u)=u$ and $x(t,\cdot)$ is order-preserving and homogeneous of degree 1, a sequence $\{\bar d(x(t_n,x_0),u)\}$ must be decreasing.

Let $\lim_{n\to\infty}\bar d(x(t_n,x_0),u)=\alpha$. If $\alpha=0$, then $x(t,x_0)\to\gamma(x_0)u$ with $\gamma(x_0)=1$. Hence we can assume that $\alpha>0$.

Then

$$\bar d\left(x(t_n,x_0),\frac{x(t_n,x_0)}{\psi(x(t_n,x_0))}\right)\to\alpha. \tag{11}$$

Since $\bar d(y,\alpha y)=|\log\alpha|$, the above implies that

$$\psi(x(t_n,x_0))\to\{e^{-\alpha},e^{\alpha}\}\quad\text{as } n\to\infty.$$

Without loss of generality we may assume that $\psi(x(t_{n_k},x_0))\to e^{\alpha}$ for some subsequence $\{t_{n_k}\}$. As $x(t,e^{\alpha}u)=e^{\alpha}u$ and the flow is order-preserving, Lemma 2.2 implies

$$\bar d(x(t,x_0),e^{\alpha}u)\leqq\bar d(x(t_{n_k},x_0),e^{\alpha}u)\quad\text{for } t\geqq t_{n_k}.$$

This implies that

$$x(t,x_0)\to e^{\alpha}u\quad\text{as } t\to\infty,$$

and that $\gamma(x_0)=e^{\alpha}$.

If $x_0, y_0 \in \operatorname{int} K$ are such that $x_0 \leqq y_0$, then

$$\gamma(x_0)u = \lim_{t\to\infty} x(t, x_0) \leqq \lim_{t\to\infty} x(t, y_0) = \gamma(y_0)u,$$

which proves that γ is order-preserving. Homogeneity of γ follows from the homogeneity of $x(t, \cdot)$.

To verify that $\gamma : \operatorname{int} K \to \mathbf{R}_+$ is a continuous function, take $x_0 \in \operatorname{int} K$ and $\varepsilon > 0$. Let $\rho > 0$ be such that $\rho|\gamma(x_0)| < \varepsilon$. Define $B = \{x \in \operatorname{int} K;\ -\rho x_0 < x - x_0 < \rho x_0\}$. Then B is open in int K and, since γ is order-preserving and homogeneous of degree 1,

$$(1-\rho)\gamma(x_0) \leqq \gamma(x) \leqq (1+\rho)\gamma(x_0)$$

for all $x \in B$.

Then we have

$$|\gamma(x) - \gamma(x_0)| \leqq \rho|\gamma(x_0)| < \varepsilon$$

for all $x \in B$, which shows that γ is continuous. □

Until now we have assumed that f has an eigenvector in the interior of K. But there are examples of maps which have eigenvectors in ∂K and not in int K. One of the difficulties in this case is that as f is only defined on K it may no longer be Fréchet differentiable at any point of the boundary of K. The Fréchet differentiability of f at u will be replaced by the existence of a Gâteaux derivative at u. In fact, many maps of interest possess Gâteaux derivative at points of ∂K. We restrict our attention to the case: $X = \mathbf{R}^n$ and K the standard cone in $\mathbf{R}^n$.

It will be convenient to introduce some notation.

If $x \in (x_1, \dots, x_n) \in \mathbf{R}^n$, then we define the support of x by

$$\operatorname{supp}(x) = \{i \in \{1, \dots, n\};\ x_i \neq 0\} \subseteq \{1, \dots, n\}.$$

If J is a subset of $\{1, \dots, n\}$ let

$$\mathbf{R}^n_J = \{x \in \mathbf{R}^n;\ \operatorname{supp}(x) \subseteq J\},$$

$$K_J = \{x \in K;\ \operatorname{supp}(x) \subseteq J\},$$

$$C_J = \{x \in \mathbf{R}^n;\ x_i \geqq 0 \text{ for all } i \notin J\},$$

$$P_J = \text{orthogonal projection onto } \mathbf{R}^n_J,$$

$$Q_J = I - P_J.$$

Let $f : K \to \mathbf{R}^n$ be a continuous map, $u \in \partial K$ and $J = \operatorname{supp}(u)$. We say that $\phi : C_J \to \mathbf{R}^n$ is a Gâteaux derivative of f at u if

$$\lim_{t\to 0^+} \frac{f(u+tx) - f(u)}{t} = \phi(x) \quad \text{for all } x \in C_J. \tag{12}$$

Note that in general ϕ is not linear.

The following properties of ϕ can be deduced from properties of f.

Lemma 2.7 [8]. *Let $f : K \to \mathbf{R}^n$ be a continuous map and $u \in \partial K$ with $J = \operatorname{supp}(u)$. Let $\phi : C_J \to \mathbf{R}^n$ be a Gâteaux derivative of f at u. If f is Lipschitz on $B_r(u) \cap K$ with a constant k, then ϕ is Lipschitz with the same constant k and the limit in* (12) *is uniform in $x \in C_J$ such that $\|x\| \leqq 1$.*

If g is a positively homogeneous map of degree 1 such that $g(K) \subseteq K$, define

$$r_K(g) = \sup\{\lambda \geqq 0;\ g(v) = \lambda v \quad \text{for some } v \in K \backslash \{0\}\},$$

and $r_K(g) = 0$ if g has no eigenvector in K.

Now we can state the next theorem.

THEOREM 2.2. *Assume that* $f: K \to \mathbf{R}^n$ *is continuous and* $f_{|\mathrm{int}\, K}$ *is locally Lipschitz. Assume that* f *generates order-preserving flow,* f *is homogeneous of degree* 1 *and that*

$$f(u) = \lambda u \quad \textit{for some } u \in \partial K \backslash \{0\} \quad \textit{with } \|u\| = 1 \quad \textit{and} \quad \lambda \in \mathbf{R}$$

and f *has no eigenvectors in* int K.

Let $J = \operatorname{supp}(u)$. *Suppose that there is an open neighborhood* U *of* u *such that* $f_{|U \cap K}$ *is Lipschitz.*

Let $h = \alpha I + f$, *for some* $\alpha \geqq 0$, *be such that* $h(K) \subseteq K$ *and that the Gâteaux derivative* ϕ *of* h *at* u *satisfies*

$$\phi(x) = \phi(P_J x) + \phi(Q_J x)$$

for any $x \in C_J$. *Finally, assume that* $A := \phi_{|\mathbf{R}^n_J}$ *is linear and irreducible and that* $r_K(Q_J \phi Q_J) < \lambda$. *Then for any* $x_0 \in \mathrm{int}\, K$, $x(t, x_0) \in \mathrm{int}\, K$ *for all* $t \geqq 0$ *and*

$$\lim_{t \to \infty} \frac{x(t, x_0)}{\|x(t, x_0)\|} = u.$$

Proof. As in the proof of Theorem 2.1, we may assume that $\alpha = 0$ and $h = f$. Because $f(0) = 0$ and $f(K) \subseteq K$, it follows that $x_i'(t, x_0) \geqq 0$. Hence $x(t, x_0) \geqq x_0 \in \mathrm{int}\, K$, and since f has a linear growth, $x(t, x_0) \in \mathrm{int}\, K$ for all $t \geqq 0$ and $x_0 \in \mathrm{int}\, K$.

Let $\psi \in K^* \backslash \{0\}$ be such that $\psi(u) = 1$. Clearly, it is enough to show that

$$\lim_{t \to \infty} \frac{x(t, x_0)}{\psi(x(t, x_0))} = u.$$

Furthermore, application of Lemma 2.3 shows that it suffices to prove the result for $x_0 \in \Sigma_0 = \{v \in \mathrm{int}\, K;\ \psi(v) = 1\}$ such that $\|x_0 - u\|$ is sufficiently small. If we write

$$y(t) = \frac{x(t, x_0)}{\psi(x(t, x_0))} = z(t) + u,$$

then the problem reduces to showing that $z(t) \to 0$ for $x_0 \in \Sigma_0$ so that $\|x_0 - u\|$ is small. A simple calculation shows that $z(t)$ satisfies

$$\begin{aligned} &z'(t) = f(z(t) + u)\psi(f(z(t) + u))(z(t) + u) := h(z(t)), \\ (13) \qquad &z(0) = x_0 - u, x_0 \in \Sigma. \end{aligned}$$

We have that

$$\lim_{s \to 0} \frac{1}{s}[h(sz) - h(0)] = \phi(z)\psi(\phi(z))u\lambda u$$

for any $z \in C_J$, and the above limit is uniform in $z \in C_J$, $\|z\| \leqq 1$.

Hence

$$h(z) = \phi(z) - \psi(\phi(z))u - \lambda z + R(z),$$

where

$$\frac{\|R(z)\|}{\|z\|} \to 0 \quad \text{as } \|z\| \to 0, z \in C_J.$$

To simplify the notation we write $P = P_J$, $Q = Q_J$, $z_1 = P_z$, and $z_2 = Qz$.

Using the assumptions about ϕ we get that

$$
\begin{aligned}
h(z) &= Az_1 + Q\phi(z_2) + P\phi(z_2) - \psi(Az_1 + \phi(z_2))u - \lambda z - R(z) \\
&= [(A-\lambda)z_1 - \psi((A-\lambda)z_1)u] + [P\phi(z_2) - \psi(\phi(z_2) - \lambda z_2)u] \\
&\quad + [Q\phi(z_2) - \lambda z_2] + R(z) \\
&= Lz_1 + g(z_2) + [Q\phi(z_2) - \lambda z_2] + R(z),
\end{aligned}
\tag{14}
$$

where $L:\mathbf{R}^n_J \to \mathbf{R}^n_J$ and $g: K_{J'} \to \mathbf{R}^n$, $(J' = \{1, \dots, n\}\backslash J)$ are defined by

$$Lz_1 = [(A-\lambda)z_1 - \psi((A-\lambda)z_1)u],$$

$$g(z_2) = [P\phi(z_2) - \psi(\phi(z_2) - \lambda z_2)u].$$

Because f is Lipschitz map, g is also Lipschitz and there is a constant $M_1 > 0$ such that

$$\|g(z_2)\| \leqq M_1\|z_2\| \quad \text{for any } z_2 \in K_{J'}.$$

The map L has only eigenvalues with real part less than zero, and so there are constants $M_2 > 0$ and $\mu > 0$ such that

$$\|e^{Lt}\| \leqq M_2\, e^{-\mu t} \quad \text{for any } t \geqq 0.$$

Consider $Q\phi$ restricted to $K_{J'}$.

Since $Q\phi(K_{J'}) \subseteq K_{J'}$ and $r_K(Q\phi Q) < \lambda$, we see that $Q\phi_{K_{J'}}$ does not have eigenvectors in $K_{J'}$ with eigenvalues $\geqq \lambda$.

Let $\eta > 0$ and $B: \mathbf{R}^n_{J'} \to \mathbf{R}^n_{J'}$ be a positive linear map. Let $T_\eta := Q\phi_{K_{J'}} + \eta B$. Then $T_\eta(K_{J'}) \subseteq \operatorname{int} K_{J'}$; applying Brouwer's fixed point theorem, we find $w_\eta \in K_{J'}$ with $\|w_\eta\| = 1$ and $\lambda_\eta > 0$ such that

$$T_\eta w_\eta = \lambda_\eta w_\eta.$$

A standard compactness argument shows that for a sufficiently small η we have that $\lambda_\eta < \lambda$.

Fix η so that $\lambda_\eta < \lambda$ and let $w := w_\eta$, $\lambda_1 := \lambda_\eta$. Define

$$(\mathbf{R}^n_{J'})_w = \{x \in \mathbf{R}^n_{J'};\ -\alpha w \leqq x \leqq \alpha w \text{ for some } \alpha > 0\},$$

and $\|x\|_w = \inf\{\alpha > 0;\ -\alpha w \leqq x \leqq \alpha w\}$. Obviously, $\mathbf{R}^n_{J'} = (\mathbf{R}^n_{J'})_w$ and the norms $\|\cdot\|$ and $\|\cdot\|_w$ are equivalent on $\mathbf{R}^n_{J'}$.

Let $x \in K_{J'}$ and $\beta > \|x\|_w$. Then $x \leqq \beta w$ and

$$
\begin{aligned}
0 \leqq Q\phi(x) \leqq \beta Q\phi(w) &= \beta Q\phi(w) + \beta\eta Bw - \beta\eta Bw \\
&= \beta T_\eta w - \beta\eta Bw = \beta\lambda_1 w - \beta\eta Bw \leqq \beta\lambda_1 w.
\end{aligned}
$$

From this we deduce that

$$\|Q\phi(x)\|_w \leqq \lambda_1\|x\|_w \quad \text{for any } x \in K_{J'}.$$

Since $y(t) = z(t) + u \geqq 0$ and $\operatorname{supp}(u) = J$, we see that $z(t) \in C_J$ for $t \geqq 0$. Now, we can write (13) as a system of two equations

$$z_1'(t) = Lz_1(t) + g(z_2(t)) + R_1(z(t)), \tag{15}$$

$$z_2'(t) = Q\phi(z_2(t)) - \lambda z_2(t) + R_2(z(t)), \tag{16}$$

where $z(t)=z_1(t)+z_2(t)=Pz(t)+Qz(t)$ and $R_1(z)=PR(z)$, $R_2(z)=QR(z)$. Observe that

$$\frac{\|R_i(z)\|}{\|z\|}\to 0 \quad \text{as } \|z\|\to 0,\ i=1,2.$$

The variation of constants formula applied to solutions of (15) and (16) gives

$$z_1(t)=e^{Lt}z_1(0)+\int_0^t e^{L(t-s)}g(z_2(s))\,ds+\int_0^t e^{L(t-s)}R_1(z(s))\,ds \tag{17}$$

and

$$z_2(t)=e^{\lambda t}z_2(0)+\int_0^t e^{\lambda(t-s)}Q\phi(z_2(s))\,ds+\int_0^t e^{\lambda(t-s)}R_2(z(s))\,ds. \tag{18}$$

These formulas together with $\|g(z_2)\leqq M_1\|z_2\|$ and $\|e^{Lt}\|\leqq M_2e^{-\mu t}$ give

$$\begin{aligned}\|z_1(t)\|\leqq M_2e^{\mu t}\|z_1(0)\|+M_1M_2\int_0^t e^{\mu(t-s)}\|z_2(s)\|\,ds\\ +M_2\int_0^t e^{\mu(t-s)}\|R_1(z(s))\|\,ds\end{aligned} \tag{19}$$

and

$$\begin{aligned}\|z_2(t)\|_w&\leqq e^{\lambda t}|z_2(0)|+\lambda_1\int_0^t e^{-\lambda(t-s)}\|z_2(s)\|_w\,ds+\int_0^t e^{\lambda(t-s)}\|R_2(z(s))\|_w\\ &\leqq e^{\lambda t}|z_2(0)|+\lambda_1\int_0^t e^{\lambda(t-s)}\|z_2(s)\|_w\,ds+\frac{1}{m_0}\int_0^t e^{\lambda(t-s)}\|R_2(z(s))\|\,ds,\end{aligned} \tag{20}$$

where m_0 is a constant so that

$$m_0\|x\|_w\leqq\|x\|<m_1\|x\|_w.$$

Let $\mu_1=\lambda-\lambda_1$, $\mu_2=\min\{\mu_1,\mu\}$, $a=m_1/m_0$ and $M_3=\max\{M_1M_2a,\ a+M_2\}$. Take $\varepsilon>0$ so that $\varepsilon<\min\{\mu/9M_3,\ 1/M_3\}$. Then there is $\delta>0$ such that

$$\|R_i(z)\|\leqq\varepsilon\|z\| \quad \text{if } \|z\|\leqq\delta,\quad i=1,2.$$

Put

$$\delta_1:=\frac{1}{M_3}\frac{\sqrt{M_3\varepsilon}}{\sqrt{M_3\varepsilon}+1}\delta.$$

We claim that if $\|z(0)\|=\|x_0-u\|\leqq\delta_1$, then $\|z(t)\|\leqq\delta$ for $t\geqq 0$. Define $t_0:=\inf\{t\geqq 0;\ \|z(t)\|=\delta\}>0$ and assume that $t_0<\infty$.

For any $t\in[0,t_0]$, we have

$$\begin{aligned}\|z_1(t)\|\leqq M_2e^{\mu_2 t}\|z_1(0)\|+M_1M_2\int_0^t e^{\mu_2(t-s)}\|z_2(s)\|\,ds\\ +\varepsilon M_2\int_0^t e^{\mu_2(t-s)}\|z(s)\|ds\end{aligned} \tag{21}$$

and

$$\begin{aligned}\|z_2(t)\|_w\leqq e^{\lambda t}|z_2(0)|+\lambda_1\int_0^t e^{-\lambda(t-s)}\|z_2(s)\|_w\,ds\\ +\frac{\varepsilon}{m_0}\int_0^t e^{\lambda(t-s)}\|z(s)\|\,ds.\end{aligned} \tag{22}$$

Taking $u(t) := e^{\lambda t}\|z_2(t)\|_w$, the second inequality is equivalent to

$$u(t) \leqq u(0) + \lambda_1 \int_0^t u(s)\, ds + \frac{\varepsilon}{m_0}\int_0^t e^{\lambda s}\|z(s)\|\, ds. \tag{23}$$

If $w(t) = \int_0^t u(s)\, ds$, the above inequality implies

$$\frac{d}{ds}(e^{\lambda_1 s}w(s)) \leqq e^{\lambda_1 s}u(0) + \frac{\varepsilon}{m_0} e^{\lambda_1 s}\int_0^s e^{\lambda \tau}\|z(\tau)\|\, d\tau,$$

which after integration on $[0, t]$ gives

$$e^{\lambda_1 t}w(t) \leqq \frac{u(0)}{\lambda_1}(1 - e^{\lambda_1 t})\frac{\varepsilon e^{\lambda_1 t}}{m_0\lambda_1}\int_0^t e^{\lambda s}\|z(s)\|\, ds + \frac{\varepsilon}{m_0\lambda_1}\int_0^t e^{(\lambda-\lambda_1)s}\|z(s)\| ds.$$

Multiplying the last inequality by $\lambda_1 e^{\lambda_1 t}$ and substituting in (22) we obtain

$$\|z_2(t)\|_w \leqq e^{\mu_2 t}|z_2(0)| + \frac{\varepsilon}{m_0}\int_0^t e^{\mu_2(t-s)}\|z(s)\|\, ds. \tag{24}$$

Since $m_0|z_2| \leqq \|z_2\| \leqq m_1|z_2|$, we get

$$\|z_2(t)\| \leqq a e^{\mu_2 t}\|z_2(0)\| + b\int_0^t e^{\mu_2(t-s)}\|z(s)\|\, ds, \tag{25}$$

where $a = m_1/m_0$ and $b = \varepsilon a$.

Now, using (25) we can estimate the second term in (21):

$$\begin{aligned}\int_0^t e^{\mu_2(t-s)}\|z_2(s)\|\, ds &\leqq a\|z_2(0)\|\int_0^t e^{\mu_2(t-s)}e^{\mu_2 s}\, ds\\ &\quad + b\int_0^t e^{\mu_2(t-s)}\left[\int_0^s e^{\mu_2(s\tau)}\|z(\tau)\|\, d\tau\right] ds\\ &= a\|z_2(0)\|e^{\mu_2 t}t + b\int_0^t (t-s)e^{-\mu_2(t-s)}\|z(s)\|\, ds.\end{aligned}$$

Thus

$$\begin{aligned}\|z_1(t)\| &\leqq (M_2 + M_1M_2at)e^{\mu_2 t}\|z_1(0)\|\\ &\quad + M_1M_2b\int_0^t e^{\mu_2(t-s)}(t-s)\|z(s)\|\, ds + \varepsilon M_2\int_0^t e^{\mu_2(t-s)}\|z(s)\|\, ds.\end{aligned} \tag{26}$$

Adding (25) and (26) we get that

$$\begin{aligned}\|z(t)\| &\leqq M_3\|z(0)\|e^{\mu_2 t}(1+t) + M_3\varepsilon\int_0^t e^{\mu_2(t-s)}\|z(s)\|\, ds\\ &\quad + M_3\varepsilon\int_0^t e^{\mu_2(t-s)}(t-s)\|z(s)\|\, ds.\end{aligned} \tag{27}$$

Let $\alpha(t) := M_3\|z(0)\|(1+t)$, $\beta(t) := e^{\mu_2 t}\|z(t)\|$ and $c := M_3\varepsilon$.

Define $N: C[0, t_0] \to C[0, t_0]$ by

$$(Nv)(t) = \int_0^t (t-s)v(s)\, ds.$$

Because $(Nv)(t) \geqq 0$ if $v(t) \geqq 0$, repeated application of N to both sides of (27) gives

$$\beta(t) \leqq \sum_{k=0}^{n} c^k(N^k\alpha)(t) + \sum_{k=0}^{n} c^{k+1}(N^k w)(t) + c^{n+1}(N^{n+1}\beta)(t) \tag{28}$$

for any $n \in N$ and $w(t) = \int_0^t \beta(s)\, ds$.

It is easy to see that

$$(N^k\alpha)(t)=\frac{1}{(2k-1)!}\int_0^t (t-s)^{2k-1}\alpha(s)\,ds$$

and

$$(N^k w)(t)=\frac{1}{(2k)!}\int_0^t (t-s)^{2k}\beta(s)\,ds.$$

As $c^{n+1}(N^{n+1}\beta)(t)\to 0$ uniformly on $[0, t_0]$, by taking a limit in (28), we deduce that

$$\begin{aligned}\beta(t)&\leqq \alpha(t)+\int_0^t \sum_{k=1}^{\infty}\frac{c^k(t-s)^{2k-1}}{(2k-1)!}\alpha(s)\,ds\\ &\quad +c\int_0^t \beta(s)\,ds+\int_0^t \sum_{k=1}^{\infty}\frac{c^{k+1}(t-s)^{2k}}{(2k)!}\beta(s)\,ds\\ &\leqq \alpha(t)+\sqrt{c}\int_0^t e^{\sqrt{c}(t-s)}\alpha(s)\,ds+c\int_0^t \beta(s)\,ds\\ &\quad +\sqrt{c}\int_0^t e^{\sqrt{c}(t-s)}\beta(s)\,ds\\ &\leqq \alpha(t)+\sqrt{c}\int_0^t e^{\sqrt{c}(t-s)}\alpha(s)\,ds+2\sqrt{c}\int_0^t e^{\sqrt{c}(t-s)}\beta(s)\,ds.\end{aligned}\tag{29}$$

If $\alpha_1(t)=e^{\sqrt{c}t}\alpha(t)+\sqrt{c}\int_0^t e^{\sqrt{c}s}\alpha(s)\,ds$, then integration by parts implies

$$\begin{aligned}\alpha_1(t)&=e^{\sqrt{c}t}\alpha(t)\int_0^t e^{-\sqrt{c}s}\alpha(s)\,ds\\ &=\alpha(0)+\int_0^t e^{\sqrt{c}s}\alpha'(s)\\ &=\alpha(0)\left(1+\frac{1}{\sqrt{c}}\right).\end{aligned}\tag{30}$$

It follows from (29) and (30) that

$$e^{\sqrt{c}t}\beta(t)\leqq \alpha(0)\left(1+\frac{1}{\sqrt{c}}\right)+2\sqrt{c}\int_0^t e^{\sqrt{c}s}\beta(s)\,ds,$$

and by Gronwall's inequality,

$$e^{\sqrt{c}t}\beta(t)\leqq \alpha(0)\left(1+\frac{1}{\sqrt{c}}\right)e^{2\sqrt{c}t}.$$

Hence

$$\beta(t)\leqq \alpha(0)\left(1+\frac{1}{\sqrt{c}}\right)e^{3\sqrt{c}t}$$

and

$$\|z(t)\|\leqq M_3\|z(0)\|\left(1+\frac{1}{\sqrt{c}}\right)e^{\mu_3 t},\tag{31}$$

where $\mu_3=3\sqrt{c}\,\mu_2<0$.

Because

$$\|z(0)\|\leqq \frac{1}{M_3}\frac{\sqrt{c}}{\sqrt{c}+1}\,\delta:=\delta_1,$$

the above inequality implies

$$\|z(t)\| \leqq \delta e^{\mu_3 t} \quad \text{for any } 0 \leqq t \leqq t_0,$$

which contradicts the definition of t_0. Hence $\|z(t)\| \leqq \delta$ for all $t \geqq 0$. Now, the same arguments show that $\|z(t)\| \leqq \delta e^{\mu_3 t}$ for all $t \geqq 0$ and this completes the proof. □

3. Applications. In this section we will illustrate applications of the results of § 2.

One of the problems in applying these results is an a priori existence of an eigenvector u of f satisfying $u \in \operatorname{int} K$ or $u \in \partial K$ and $r_K(Q\phi Q) < \lambda$. Even for a simple-looking map these questions can be difficult. We will define classes of maps $\mathcal{M}$, $\mathcal{M}_-$, $\mathcal{M}_+$ which were rigorously analyzed by Nussbaum in [7] and [8]. To keep this section self-contained we will quote some of his results.

Suppose that K is the standard cone in $\mathbf{R}^n$. If $r \in \mathbf{R}$, and $\sigma = (\sigma_1, \ldots, \sigma_n)$ is a probability vector, i.e., $\sigma_i \geqq 0$, $i = 1, \ldots, n$ and $\sum_{i=1}^n \sigma_i = 1$, define for $x \in \operatorname{int} K$

$$M_{r\sigma}(x) = \left[\sum_{i=1}^n \sigma_i x_i^r \right]^{1/r} \quad \text{if } r \neq 0$$

and

$$M_{0\sigma}(x) = \prod_{j=1}^n x_j^{\sigma} = \lim_{r \to 0} M_{r\sigma}(x) \quad \text{for } x \in \operatorname{int} K.$$

Each of the maps $M_{r\sigma}$ is obviously C^∞ on $\operatorname{int} K$ and extends continuously to K. For each i, let Γ_i be a finite collection of ordered pairs (r, σ), $r \in \mathbf{R}$ and σ a probability vector.

For each pair $(r, \sigma) \in \Gamma_i$ assume that $c_{ir\sigma}$ is a positive number. Let

$$f_i(x) = \sum_{(r,\sigma) \in \Gamma_i} c_{ir\sigma} M_{r\sigma}(x) \tag{32}$$

and $f(x) = (f_1(x), \ldots, f_n(x))$. Of course, $f: \operatorname{int} K \to \operatorname{int} K$ and f can be continuously extended to K.

If $f: \operatorname{int} K \to \operatorname{int} K$ is a map such that each of its components can be written as in (32), we say that $f \in M$.

If $f \in M$ and each of the components f_i has the form (32) with $r \geqq 0$ for all $(r, \sigma) \in \Gamma_i$ we write $f \in M_+$ and if each component f_i can be written as in (32) with $r < 0$ for all $(r, \sigma) \in \Gamma_i$, then $f \in M_-$.

Note that the sets Γ_i are not uniquely determined by f_i, for example $(x^r)^{1/r} = x$. By $\mathcal{M}(\mathcal{M}_+, \mathcal{M}_-)$ we denote the smallest sets of maps $f: \operatorname{int} K \to \operatorname{int} K$ such that $M \subseteq \mathcal{M}(M_+ \subseteq \mathcal{M}_+, M_- \subseteq \mathcal{M}_-)$ and $\mathcal{M}(\mathcal{M}_+, \mathcal{M}_-)$ is closed under addition of functions, composition of functions and multiplication by positive numbers. It is known (see [8]) that if $f \in \mathcal{M}(\mathcal{M}_+, \mathcal{M}_-)$, then $f_{|\operatorname{int} K}$ is C^∞ and extends continuously to K and is order-preserving. Hence, by Remark 2.1, the flow of (1) with f in one of these classes is automatically order-preserving.

Definition 3.1. If $g: \operatorname{int} K \to K$ is continuous, order-preserving and homogeneous of degree 1 map, then the nonnegative $n \times n$ matrix $A = (a_{ij})$ is called an "incidence matrix for g" if whenever $a_{ij} > 0$, there is a positive c and a probability vector σ such that the jth component of σ, σ_j, is positive and

$$g_i(x) \geqq cx^{\sigma} \quad \text{for all } x \in \operatorname{int} K.$$

For maps with incidence matrices Nussbaum proves the following.

Theorem 3.1. *Assume that $g: \operatorname{int} K \to \operatorname{int} K$ is C^1, g extends continuously to K, g is positively homogeneous of degree 1 and order-preserving. If g has an incidence matrix A which is irreducible, then g has an eigenvector $u \in \operatorname{int} K$ with $\|u\| = 1$.*

Many maps in $\mathcal{M}_+$ have irreducible incidence matrices. If $g \in \mathcal{M}_+$, then $D_x g(x) = A$ is an incidence matrix for g and $D_x g(x)$, $D_x g(y)$ satisfy

$$\alpha D_x g(x) \leqq D_x g(y) \leqq \beta D_x g(x) \tag{33}$$

for some $\alpha, \beta > 0$, depending on $x, y \in \operatorname{int} K$.

Having this the following corollary is an immediate consequence of Theorem 2.1.

COROLLARY 3.1. *Let $f: K \to \mathbf{R}^n$ be such that for some $\alpha \geqq 0$ the map $g: K \to K$ defined by $g(x) = \alpha x + f(x)$ is in the class $\mathcal{M}_+$. Assume that there is $x \in \operatorname{int} K$ such that $D_x g(x)$ is irreducible.*

Then g has an eigenvector $u \in \operatorname{int} K$ such that $\|u\| = 1$, and for any $x_0 \in \operatorname{int} K$ we have that

$$\frac{x(t, x_0)}{\|x(t, x_0)\|} \to u, \qquad t \to \infty.$$

Proof. That g has an eigenvector $u \in \operatorname{int} K$ follows from Theorem 3.1. Then, since $D_x g(x)$ is irreducible and (33) holds, it follows that $D_x g(u)$ is also irreducible. By Theorem 2.1, if $y(t)$ is a solution of $y'(t) = g(y(t))$, we have

$$\frac{y(t, x_0)}{\|y(t, x_0)\|} \to u \quad \text{for any } x_0 \in \operatorname{int} K.$$

But $y(t, x_0) = e^{\alpha t} x(t, x_0)$, $x'(t) = f(x(t))$ and the conclusion follows. □

The existence of eigenvectors in $\operatorname{int} K$ for maps in the class $\mathcal{M}_-$ is more subtle.

Let A be a subset of M_- such that if $g \in A$, then for any two elements (r, σ) and (r', σ') of Γ_i, $1 \leqq i \leqq n$, we have that σ and σ' are comparable.

We denote by $\mathcal{A}$ the smallest set of functions $g: \operatorname{int} K \to \operatorname{int} K$ which is closed under composition and contains A. It is known that if $g \in \mathcal{A}$, then $g: \operatorname{int} K \to \operatorname{int} K$, g is C^1, order-preserving, and homogeneous of degree 1. Moreover, if $D_x g(x)$ is irreducible for some $x \in \operatorname{int} K$, then g has an eigenvector u in $\operatorname{int} K$.

COROLLARY 3.2. *Let $f: \operatorname{int} K \to \mathbf{R}^n$ be such that $g = \alpha I + f$ belongs to $\mathcal{A}$. If for some $x \in \operatorname{int} K$, $D_x g(x)$ is irreducible, then*

$$\lim_{t \to \infty} \frac{x(t, x_0)}{\|x(t, x_0)\|} = u \qquad \text{for each } x_0 \in \operatorname{int} K,$$

where u is an eigenvector of g in int K.

Proof. By the previous remarks the proof is similar to the proof of Corollary 3.1. □

Now we will give a result concerning convergence to a boundary point of ∂K for maps in $\mathcal{M}_-$. We have the following.

PROPOSITION 3.1. *Let $g \in \mathcal{M}_-$. Then g is Lipschitz on K. If $u \in \partial K - \{0\}$, then for all $x \in C_J$,*

$$\lim_{t \to t^+} t^{-1}[g(u + tx) - g(u)] \equiv \phi(x)$$

exists and $\phi \in \mathcal{M}_-$.

Moreover, the above limit is uniform for $u \in C_J$ such that $\|u\| \leqq 1$. For all $u \in C_J$

$$\phi(u) = \phi(P_J u) + \phi(Q_J u)$$

and $\phi_{|\mathbf{R}^n_J}$ is linear. If $D_x g(x)$ is irreducible at $x \in$ int K and $g(u) = \lambda u$ for some $u \in \partial K \backslash \{0\}$ with supp $(u) = J$ *and*

$$r_K(Q_J \phi Q_J) \leqq \lambda,$$

then g has no eigenvector in int K.

Using the above proposition and Theorem 2.2 we obtain the following.

COROLLARY 3.3. *Assume that* $f\colon K\to \mathbf{R}$ *and* g *is defined by* $g(x)=\alpha x+f(x)$, $\alpha \geqq 0$. *Assume that* $D_xg(x)$ *is irreducible for some* $x\in \operatorname{int} K$ *and that* $g(u)=\lambda u$ *for some* $u\in \partial K\backslash\{0\}$ *with* $\operatorname{supp}(u)=J$ *and* $\|u\|=1$. *Moreover, assume that*

$$r_K(Q_J\phi Q_J)<\lambda.$$

Then for any $\xi_0\in \operatorname{int} K$,

$$\lim_{t\to\infty}\frac{x(t,\xi_0)}{\|x(t,\xi_0)\|}=u.$$

Proof. The proof follows from Theorem 2.2 and the previous proposition. □

We will close this section by considering a very special example. Let $f\colon K\to \mathbf{R}^4$ by

$$\begin{aligned}
f_1(x)&=\alpha_1x_1+\beta_1\theta(x_1,x_2)+\gamma_1\theta(x_1,x_4)+\delta_1\theta(x_2,x_3),\\
f_2(x)&=\alpha_2x_2+\beta_2\theta(x_1,x_2)+\gamma_2\theta(x_1,x_4)+\delta_2\theta(x_2,x_3),\\
f_3(x)&=\alpha_3x_3+\beta_3\theta(x_3,x_4)+\gamma_3\theta(x_1,x_4)+\delta_3\theta(x_2,x_3),\\
f_4(x)&=\alpha_4x_4+\beta_4\theta(x_3,x_4)+\gamma_4\theta(x_1,x_4)+\delta_4\theta(x_2,x_3).
\end{aligned}\tag{34}$$

Here θ stands for the harmonic mean:

$$\theta(s,t)=\frac{st}{s+t}=\frac{1}{2}\left[\frac{1}{2}s^{-1}+\frac{1}{2}t^{-1}\right]^{-1}.$$

The coefficients β_i, γ_i, δ_i satisfy

(H.1) β_i, γ_i, δ_i are nonnegative for $1\leqq i\leqq 4$;

(H.2) δ_1, γ_2, γ_3, δ_4 are strictly greater than zero;

(H.3) $\alpha_1<\alpha_2+\beta_2$, $\alpha_2<\alpha_1+\beta_1$, $\alpha_3<\alpha_4+\beta_4$, $\alpha_4<\alpha_3+\beta_3$.

It is easy to see that if $\alpha>|\alpha_i|$ for $1\leqq i\leqq 4$, then $g=\alpha I+f$ is in the class $\mathcal{M}_-$.

This four-dimensional map has been introduced by Schoen in his studies of a model from the population biology.

The following is known about f; for the proof see [8].

Let $a_1, a_2\in \mathbf{R}$, and $b_i, c_i\geqq 0$ for $i=1,2$, and assume that at least one of b_1, c_1, b_2, or c_2 is positive. Consider the equation

$$a_1+b_1t(1+t)^{-1}c_1t=a_2+b_2(1+t)^{-1}+c_2t^{-1}.$$

If this equation has a positive solution, then it is unique, say τ. Define

$$\sigma(a_1,b_1,c_1,a_2,b_2,c_2)=a_1+b_1\tau(1+\tau)^{-1}+c_1\tau.$$

If $c_1=0$ and $c_2>0$ and $a_2\geqq a_1+b_1$, let

$$\sigma(a_1,b_1,c_1,a_2,b_2,c_2)=a_2.$$

If $c_1>0$ and $c_2=0$ and $a_1\geqq a_2+b_2$, let

$$\sigma(a_1,b_1,c_1,a_2,b_2,c_2)=a_1.$$

If $c_1=c_2=0$ and $a_1\geqq a_2+b_2$ or $a_2\geqq a_1+b_1$, define

$$\sigma(a_1,b_1,c_1,a_2,b_2,c_2)=\max(a_1,a_2).$$

If

$$\sigma(\alpha_1+\gamma_1,\beta_1,\delta_1,\alpha_2+\delta_2,\beta_2,\gamma_2)\leqq\sigma(\alpha_3,\beta_3,0,\alpha_4,\beta_4,0)\tag{35}$$

or

$$(36) \qquad \sigma(\alpha_3+\delta_3, \beta_3, \gamma_3, \alpha_4+\gamma_4, \beta_4, \delta_4) \leqq \sigma(\alpha_1, \beta_1, 0, \alpha_2, \beta_2, 0),$$

then f has no eigenvector in int K.

If neither inequality is satisfied, then f has a unique eigenvector in int K. Having this we can state the following.

COROLLARY 3.4. *Let f be defined by* (34) *and let f satisfy* (H.1), (H.2), *and* (H.3). *If none of the inequalities* (35) *and* (36) *are satisfied, then*

$$\frac{x(t, x_0)}{\|x(t, x_0)\|} \to u \quad \textit{for all } x_0 \in \text{int } K,$$

where u is the unique normalized eigenvector of f in int K.

Proof. It is easy to see that if $\alpha > \sum_{i=1}^4 |\alpha_i|$ and $g = \alpha I + f$ is in $\mathcal{M}_-$, then $D_x g(x)$ is irreducible at each $x \in \text{int } K$. Thus g has a unique eigenvector u in int K, and the result follows from Corollary 3.2. □

If β_1 and β_2 are positive and (36) is satisfied with strict inequality, then f has a unique normalized eigenvector $u = (u_1, u_2, 0, 0)$, $u_i > 0$.

If β_3 and β_4 are positive and (35) is satisfied with strict inequality, then f has a unique normalized eigenvector $v = (0, 0, v_3, v_4)$, $v_i > 0$. Hence we can state the following.

COROLLARY 3.5. *If f is as in* (34) *and either* (35) *or* (36) *is satisfied with strict inequality, then we have*

$$\lim_{t\to\infty} \frac{x(t, \xi_0)}{\|x(t, \xi_0)\|} = w \quad \textit{for any } \xi_0 \in \text{int } K,$$

where $w = (w_1, w_2, 0, 0)$ if $\beta_1, \beta_2 > 0$ and (36) *holds with strict inequality, or $w = (0, 0, w_3, w_4)$ if $\beta_3, \beta_4 > 0$ and* (35) *holds with strict inequality.*

Proof. The proof follows from Corollary 3.3. □

Acknowledgment. I would like to thank Professor Roger Nussbaum for suggesting this problem and for helpful discussions.

REFERENCES

[1] G. BIRKHOFF AND L. KOTLIN, *Essentially positive systems of linear differential equations*, Bull. Amer. Math. Soc., 71 (1965), pp. 771-772.

[2] P. BUSHELL, *Hilbert's metric and positive contraction mappings in Banach space*, Arch. Rational Mech. Anal., 52 (1973), pp. 330-338.

[3] ———, *The Caley-Hilbert metric and positive operators*, Linear Algebra Appl., 84 (1986), pp. 271-280.

[4] M. HIRSCH, *The dynamical systems approach to differential equations*, Bull. Amer. Math. Soc., 11 (1984), pp. 1-64.

[5] M. A. KRASNOSELSKII, *Translation Along Trajectories of Differential Equations*, 19, Transl. Math. Monographs, American Mathematical Society, Providence, RI, 1968.

[6] R. D. NUSSBAUM, *Eigenvectors of nonlinear positive operators and the linear Krein-Rutman theorem*, in Fixed Point Theory, Springer Lecture Notes in Math., 886, Springer, New York, 1981, pp. 309-331.

[7] ———, *Hilbert's Projective Metric and Iterated Nonlinear Maps, Part* I, Mem. Amer. Math. Soc., 75 (1988).

[8] ———, *Hilbert's Projective Metric and Iterated Nonlinear Maps, Part* II, Mem. Amer. Math. Soc., 79 (1989).

[9] P. A. SAMUELSON, *Generalizing Fisher's "reproductive value" nonlinear homogeneous systems*, Proc. Nat. Acad. Sci. U.S.A., 74 (1977), pp. 5772-5775.

[10] H. SCHAEFER, *Topological Vector Spaces*, Springer-Verlag, New York, 1971.

[11] T. SCHANBACHER, *Asymptotic behavior of positive semigroups*, Math. Zeitschrift, 1195 (1987), pp. 481-485.

[12] R. SCHOEN, *The two-sex multi-ethnic stable population model*, Theoret. Pop. Biol., 129 (1986), pp. 343-364.

[13] H. R. THIEME, *Asymptotic proportionality (weak ergodicity) and conditional asymptotic equality of solutions to time-heterogeneous sublinear difference and differential equations*, J. Differential Equations, 73 (1988), pp. 237-268.

[14] A. C. THOMPSON, *On Certain Contraction Mappings in a Partially Ordered Vector Space*, Proc. Amer. Math. Soc., 1963, pp. 438-443.

SIAM J. MATH. ANAL.
Vol. 24, No. 3, pp. 703-711, May 1993

010

EXPLICIT FORMULA FOR THE BEST LINEAR PREDICTOR OF PERIODICALLY CORRELATED SEQUENCES*

A. G. MIAMEE†

Abstract. Explicit formulas expressing the best linear predictor and prediction error of a periodically correlated sequence in terms of a related multivariate stationary sequence are obtained. Algorithms for determining this best linear predictor and this predictor error are also outlined.

Key words. periodically correlated sequences, best linear predictor, prediction error, multivariate stationary sequences

AMS subject classifications. 60G10, 60G25

1. Introduction. A sequence X_n, $n \in Z$ of elements of a Hilbert space H is called a *periodically correlated sequence* (PCS) if there exists an integer $T > 0$ such that for any τ, $n \in Z$,

$$(X_{n+\tau}, X_n) = (X_{n+T+\tau}, X_{n+T}).$$

The *smallest* such T is called its period. (If $T = 1$ the sequence is called *stationary.*) For a PCS X_n, its *correlation function*

$$R(n, \tau) = (X_{n+\tau}, X_n)$$

is periodic in n with period T and, therefore, has the representation [1]

$$R(n, \tau) = \sum_{k=0}^{T-1} R_k(\tau) \exp(2\pi ikn/T). \tag{1}$$

For convenience we complete the definition of the functions $R_k(\tau)$, $k = 0, \dots, T-1$ to all integers using $R_k(\tau) = R_{k+T}(\tau)$.

There is a very close tie between the class of PCS's and that of multivariate stationary sequences (SS). For example it is well known and easy to check that: *A sequence X_n is periodically correlated if and only if the T-variate sequence $\underline{Y}_n = (X_{nT}, X_{nT+1}, \dots, X_{nT+T-1})^T$ constructed from consecutive blocks of length T from the process X_n is stationary.* Another such close tie is the subject of the following result proved in [1].

THEOREM 1.1 (Gladyshev). *Function* (1) *is the correlation function of some* PCS *if and only if the matrix-valued function*

$$\underline{R}(\tau) = (R_{kk'}(\tau))_{k,k'=0}^{T-1}$$

with

$$R_{kk'}(\tau) = R_{k'-k}(\tau) \exp(2\pi ik\tau/T), \tag{2}$$

is the matrical correlation function of some T-variate SS $\underline{Z}_n$.

Using this theorem and the well-known spectral representation of multivariate SS's we can show [9], [10]

$$R_k(\tau) = \int_0^{2\pi} \exp(-i\tau\lambda)\, dF_k(\lambda), \tag{3}$$

* Received by the editors December 2, 1991; accepted for publication (in revised form) August 6, 1992. This research was supported by U.S. Army Research Office grant DAAL 03-91-G-0238 and Office of Naval Research grant N00014-89-J-1824.

† Department of Mathematics, Hampton University, Hampton, Virginia 23668.

where dF_k's are complex-valued measures defined on $[0, 2\pi)$. Identifying $[0, 2\pi)$ with the unit circle in the standard way we can extend the definition of the measures dF_k beyond the interval $[0, 2\pi)$.

We note from formulas (1) and (2) that the periodically correlated sequence X_n must be harmonizable, i.e.,

$$X_n = \int_0^{2\pi} e^{in\lambda}\, dZ(\lambda),$$

where $dZ(\lambda)$ is an H-valued measure satisfying

$$(Z(\Lambda'), Z(\Lambda))_H = \sum_{k=-T+1}^{T-1} F_k(\Lambda' \cap (\Lambda - 2\pi k/T)),$$

($\Lambda - a$ stands for the set of all points $\lambda - a$, with $\lambda \in \Lambda$). In other words the spectrum of the PCS X_n is concentrated on $2T-1$ straight line segments $\lambda - \mu = 2\pi k/T$, $k = -T+1, \dots, T-1$ contained in the square $[0, 2\pi) \times [0, 2\pi)$. For more information on PCS's, see [1], [3], [4], [7].

It turns out that this second associated multivariate SS $\underline{Z}_n$ is more useful for prediction purposes than the first one. This is because of the fact that it captures more prediction properties of the original PCS X_n (see next section). However, so far no explicit representation of this $\underline{Z}_n$ process in terms of the original PCS X_n (in time domain) is available in the literature, and this is needed for our prediction purposes here. In § 3 we will obtain explicit expressions of the original PCS X_n and its associated multivariate SS $\underline{Z}_n$ (of Theorem 1.1) in terms of each other. These explicit expressions play an important role in finding our formulas for the best linear predictor and prediction error in § 4. These expressions seem to be equally important in studying other prediction problems concerning the PCS's.

2. Preliminaries. Let H be a Hilbert Space whose inner product is denoted by $(,)$, and let T be a positive integer. The direct sum $H \oplus, \dots, \oplus H$ of T copies of H is defined to be the Cartesian product of T copies of H. However, the elements of $H \oplus, \dots, \oplus \mathrm{H}$ are denoted by $X_0 \oplus, \dots, \oplus X_{T-1}$ rather than by $(X_0, X_1, \dots, X_{T-1})$, which is saved for the space H^T to be defined shortly. It is well known that $H \oplus, \dots, \oplus H$ with the usual addition and scalar multiplication becomes a Hilbert space when equipped with the *Euclidean inner product*

$$\langle X, Y \rangle = \sum_{j=0}^{T-1} (X_j, Y_j)$$

of $X = X_0 \oplus, \dots, \oplus X_{T-1}$ and $Y = Y_0 \oplus, \dots, \oplus Y_{T-1}$. Throughout this paper the direct sum $H \oplus, \dots, \oplus H$ will be denoted by K.

Following [6], H^T denotes the *Cartesian product* of H with itself T times, i.e., the set of all column vectors $\underline{X} = (X^0, \dots, X^{T-1})^T$ with $X^i \in H$ for all $i = 0, \dots, T-1$. We endow the space H^T with a Gramian structure: For $\underline{X}$ and $\underline{Y}$ in H^T their *Gramian* denoted by $((\underline{X}, \underline{Y}))$ is given by

$$((\underline{X}, \underline{Y})) = [(X^i, Y^j)]_{i,j=0}^{T-1}.$$

We can easily verify that

$$((\underline{X}, \underline{X})) \geqq \underline{0}; \qquad (\underline{X}, \underline{X}) = \underline{0} \Leftrightarrow \underline{X} = \underline{0};$$

$$\left(\left(\sum_{k=1}^{m} \underline{A}_k \underline{X}_k, \sum_{l=1}^{n} \underline{B}_l \underline{Y}_l\right)\right) = \sum_{k=1}^{m} \sum_{l=1}^{n} \underline{A}_k ((\underline{X}_k, \underline{Y}_l)) \underline{B}_l^*$$

for any $\underline{X}, \underline{X}_k, \underline{Y}_l \in H^T$ and any $T \times T$ matrices, $\underline{A}_k, \underline{B}_l$. We say that $\underline{X}$ is *orthogonal* to $\underline{Y}$ in H^T if $((\underline{X}, \underline{Y})) = \underline{0}$. A *closed subset* $\underline{M}$ of H^T is called a *subspace* if $\underline{A}\underline{X} + \underline{B}\underline{Y} \in \underline{M}$ whenever $\underline{X}, \underline{Y} \in \underline{M}$ and $\underline{A}, \underline{B}$ are $T \times T$ matrices. It is interesting and easy to see that $\underline{M}$ is a subspace of H^T if there exists a subspace M of H such that $\underline{M} = M^T$ [9], [10]. Let's denote the *orthogonal projection* of $X \in H$ on a subspace M of H by $(X \mid M)$. Given a vector $\underline{X} = (X^0, \ldots, X^{T-1}) \in H^T$, its projection on a subspace $\underline{M} = M^T$, denoted by $(\underline{X} \mid \underline{M})$, is the vector whose ith component is $(X^i \mid M)$ for each $i = 0, \ldots, T-1$. It is easy to see that $(\underline{X} \mid \underline{M})$ is the unique vector in $\underline{M}$ such that $\underline{X} - (\underline{X} \mid \underline{M}) \perp \underline{Y}$ for all $\underline{Y} \in \underline{M}$. For any set of vectors $\{\underline{X}_j : j \in J\}$ in H^T, the span closure $\overline{\text{sp}}\,\{\underline{X}_j : j \in J\}$ is the smallest closed subspace of H^T containing all the linear combinations of $\underline{X}_j$'s formed with *$T \times T$ matrices* as coefficients.

A sequence $\underline{X}_n$ in H^T is called a *stationary sequence* if the Gramian $((\underline{X}_m, \underline{X}_n))$ depends *only* on $m - n$.

For a sequence X_n in H, we define its *past-present subspace* at time n by

$$H(X; n) = \overline{\text{sp}}\,\{X_k : k \leqq n\}$$

and its *remote past subspace* by

$$H(X; -\infty) = \bigcap_n H(X; n).$$

The past-present subspace $\underline{H}(\underline{X}; n)$ and the remote past $\underline{H}(\underline{X}, -\infty)$ of a sequence $\underline{X}_n$ in H^T are similarly defined to be the subspaces

$$\underline{H}(\underline{X}; n) = \overline{\text{sp}}\,\{\underline{X}_k : k \leqq n\},$$
$$\underline{H}(\underline{X}; -\infty) = \bigcap_n \underline{H}(\underline{X}; n)$$

of H^T.

For a sequence $\underline{X}_n$ in H^T we can also define the subspace $H(\underline{X}; n)$ of H by

$$H(\underline{X}; n) = \overline{\text{sp}}\,\{X_k^j : k \leqq n, 0 \leqq j \leqq T-1\}.$$

A sequence $X_n(\underline{X}_n)$ in $H(H^T)$ is called *nondeterministic* if $X_n(\underline{X}_n)$ does not belong to $H(\underline{X}; n-1)(\underline{H}(\underline{X}; n-1))$ for every n, and it is called *purely nondeterminstic* if $H(X; -\infty) = 0$ $(\underline{H}(\underline{X}; -\infty) = \underline{0})$.

For a nondeterministic sequence $X_n(\underline{X}_n)$ in $H(H^T)$ the *best linear predictor* $X_{n,\nu}(\underline{X}_{n,\nu})$ of a future value $X_{n+\nu}(\underline{X}_{n+\nu})$, given its past-present $H(X; n)(\underline{H}(\underline{X}; n))$ is its projection

$$X_{n,\nu} = (X_{n+\nu} \mid H(X; n))(\underline{X}_{n,\nu} = (\underline{X}_{n+\nu} \mid \underline{H}(\underline{X}; n)))$$

on $H(X; n)(\underline{H}(\underline{X}; n))$.

3. Explicit expression of PCS X_n and its associate multivariate SS $\underline{Z}_n$ in terms of each other. In this section we will give some explicit expressions between a PCS X_n in H and its associated SS $\underline{Z}_n$ (mentioned in § 1) in K^T. Note that Gladyshev's Theorem 1.1 gives *only* the relation between the *correlation functions* of X_n and $\underline{Z}_n$. However, in order for us to use this theorem for the prediction of a PCS, we should know the relation between these processes *themselves.*

We start with the following theorem.

THEOREM 3.1. *A sequence X_n in H is periodically correlated with period T if and only if the sequence $\underline{Z}_n = [Z_n^k]_{k=0}^{T-1}$ in K^T with*

$$Z_n^k = \bigoplus_{j=0}^{T-1} X_{n+j} \exp\left(2\pi i k(n+j)/T\right)$$

is stationary.

Proof. If part. Assuming $\underline{Z}_n$ is a SS in K^T then its 0th component Z_n^0 is a stationary sequence in K. So, for each $m, n \in Z$ we have

$$\langle Z_m^0, Z_n^0 \rangle = \langle Z_{m+1}^0, Z_{n+1}^0 \rangle.$$

Expanding each side and multiplying through by T we get

$$\sum_{j=0}^{T-1} (X_{m+j}, X_{n+j}) = \sum_{j=0}^{T-1} (X_{m+1+j}, X_{n+1+j}) \quad \text{for all } m, n \in Z.$$

Eliminating like terms from both sides of this equality, we get

$$(X_m, X_n) = (X_{m+T}, X_{n+T}) \quad \text{for all } m, n \in Z.$$

This shows that X_n is a PCS with Period T.

Only if part. For any $m, n \in Z$ and any $0 \leqq k, k' \leqq T-1$ we can write

$$\begin{aligned} T\langle Z_m^k, Z_n^{k'} \rangle &= \sum_{j=0}^{T-1} (X_{m+j}, X_{n+j} \exp(2\pi i(km + kj - k'n - k'j)/T) \\ &= (X_m, X_n) \exp(2\pi i(km - k'n)/T) \\ &\quad + \sum_{j=1}^{T-1} (X_{m+j}, X_{n+j}) \exp(2\pi i(km + kj - k'n - k'j)/T) \\ &= \sum_{j=1}^{T-1} (X_{m+j}, X_{n+j}) \exp(2\pi i(km + kj - k'n - k'j)/T) \\ &\quad + (X_{m+T}, X_{n+T}) \exp(2\pi i(km + kT - k'n - k'T)/T) \\ &= \sum_{j=1}^{T} (X_{m+j}, X_{n+j}) \exp(2\pi i(km + kj - k'n - k'j)/T). \end{aligned}$$

The third equality follows from the periodicity assumption on X_n. Letting $J = j-1$, we get

$$\begin{aligned} T\langle Z_m^k, Z_n^{k'} \rangle &= \sum_{J=0}^{T-1} (X_{m+J+1}, X_{n+J+1}) \exp(2\pi i(km + k(J+1) - k'n - k'(J+1))/T) \\ &= \sum_{J=0}^{T-1} (X_{(m+1)+J}, X_{(n+1)+J}) \exp(2\pi i[k(m+1) + kJ - k'(n+1) - k'J]/T) \\ &= T\langle Z_{m+1}^k, Z_{n+1}^{k'} \rangle, \end{aligned}$$

which shows that $\underline{X}_n$ is stationary and completes the proof. □

Now, assuming that X_n is a PCS with correlation function (1) we want to find the matricial correlation function of its associated SS $\underline{Z}_n$ introduced in the last theorem.

For any $\tau \in Z$ and $0 \leqq k, k' \leqq T-1$ the matricial correlation function of $\underline{Z}_n$ is given by

$$\begin{aligned} R_{kk'}(\tau) = \langle Z_\tau^k, Z_0^{k'} \rangle &= \frac{1}{T} \sum_{j=0}^{T-1} (X_{\tau+j}, X_j) \exp(2\pi i(k\tau + kj - k'j)/T) \\ &= \frac{1}{T} \sum_{j=0}^{T-1} R(j, \tau) \exp(2\pi i(kj + k\tau - k'j)/T) \\ &= \frac{1}{T} \exp(2\pi i k\tau/T) \sum_{j=0}^{T-1} R(j, \tau) \exp(2\pi i(k-k')j/T). \end{aligned}$$

By inverting the Fourier sum (1) we get

$$R_k(\tau)=\frac{1}{T}\sum_{n=0}^{T-1} R(n,\tau)\exp(-2\pi ikn/T).$$

Using this we can continue to write

$$R_{kk'}(\tau)=\exp(2\pi ik\tau/T)R_{k'-k}(\tau).$$

This shows that the process $\underline{Z}_n$ introduced in Theorem 3.1 has actually the same matricial correlation suggested in Gladyshev's Theorem 1.1. So, we have obtained an explicit expression for the SS $\underline{Z}_n$ in terms of the original PCS X_n.

The following proposition helps us to express the PCS X_n in terms of its associated SS $\underline{Z}_n$. In order to do that, however, we must identify H with the subspace $H\oplus 0\oplus\cdots\oplus 0$ of K.

PROPOSITION 3.2. *If the matricial correlation function* $(R_{kk'}(\tau))_{k,k'=0}^{T-1}$ *of a* SS $\underline{Z}_n=[Z_n^k]_{k=0}^{T-1}$ *in* K^T *and the correlation function* $R(n,\tau)$ *are related as in the Theorem* 1.1, *then the sequence*

$$X_n=\sum_{k=0}^{T-1}\frac{1}{\sqrt{T}}Z_n^k\exp(-2\pi ikn/T)$$

is a PCS *in* K *with correlation function* $R(n,\tau)$.

Proof. We can write

$$\begin{aligned}(X_{n+\tau},X_n)&=\left(\frac{1}{\sqrt{T}}\sum_{k=0}^{T-1}Z_{n+\tau}^k\exp\left(-2\pi i((n+\tau)k/T),\frac{1}{\sqrt{T}}\sum_{k'=0}^{T-1}Z_n^{k'}\exp(-2\pi ink'/T)\right)\right)\\&=\frac{1}{T}\sum_{k=0}^{T-1}\sum_{k'=0}^{T-1}(Z_{n+\tau}^k,Z_n^{k'})\exp(-2\pi ik\tau/T)\exp(2\pi in(k'-k)/T)\\&=\frac{1}{T}\sum_{k=0}^{T-1}\sum_{k'=0}^{T-1}R_{kk'}(\tau)\exp(-2\pi ik\tau/T)\exp(2\pi in(k'-k)/T)\\&=\frac{1}{T}\sum_{k=0}^{T-1}\sum_{k'=0}^{T-1}R_{k'-k}(\tau)\exp(2\pi in(k'-k)/T)\\&=\frac{1}{T}\sum_{j=0}^{T-1}TR_j(\tau)\exp(2\pi inj/T)=R(n,\tau).\end{aligned}$$

The fourth equality follows from (2) and the last one from (1). □

We close this section with the following Lemma.

LEMMA 3.3. *Let* X_n *be a* PCS *whose correlation function* $R(n,\tau)$ *and spectral measures* $dF_k(\lambda)$ *satisfy* (1) *and* (3), *respectively. Then, the matricial spectral measure* $d\underline{F}_x=(dF_{kk'})_{k,k'=0}^{T-1}$ *defined by*

(4) $$F_{kk'}(\Lambda)=F_{k'-k}(\Lambda+2\pi k/T)\quad \textit{for each Borel subset } \Lambda \textit{ of } [0,2\pi),$$

is the spectral measure of its associated SS $\underline{Z}_n$.

Proof. From (2) and (3) we can write

$$\begin{aligned}(Z_\tau^k,Z_0^{k'})&=\exp(2\pi ik\tau/T)\int_0^{2\pi}\exp(-i\lambda\tau)\,dF_{k'-k}(\lambda)\\&=\int_0^{2\pi}\exp(-i\tau(\lambda-2\pi ik/T)\,dF_{k'-k}(\lambda)).\end{aligned}$$

The proof is now immediate. □

4. Explicit formulas for the predictor and predictor error of a PCS. The problem of finding an algorithm for determining the best linear predictor is important in applications of stochastic processes. In this section we will give formulas expressing the best linear predictor and prediction error of PCS X_n in terms of the predictor and prediction error matrix of its associated multivariate SS $\underline{Z}_n$ given in the last sections. Using this, along with the available algorithms for determining the predictor of multivariate SS's, we then will outline algorithms for finding the predictor and prediction error of X_n.

The following theorem gives formulas for expressing the ν-step predictor $X_{n,\nu}$ of a PCS X_n in terms of the predictor of its associated SS $\underline{Z}_n$ given in Theorems 1.1 and 3.1. We would like to point out, however, that our SS $\underline{Z}_n$ here is quite different from the SS Z_n used in [7]. In fact, Z_n is just the 0th component Z_n^0 of the SS $\underline{Z}_n$. This new SS $\underline{Z}_n$ captures more properties of X_n, and hence seems to be more useful for prediction purposes.

THEOREM 4.1. *If the ν-step predictor of the associated* SS $\underline{Z}_n$ *of a* PCS X_n *has an autoregressive representation*

$$\underline{Z}_{n,\nu} = \sum_{k=0}^{\infty} \underline{A}_k^{\nu} \underline{Z}_{n-k}, \tag{5}$$

then the best linear predictor $X_{n,\nu}$ of $X_{n+\nu}$ based on its past-present at n can be expressed as

$$X_{n,\nu} = \frac{1}{\sqrt{T}} \sum_{k=0}^{\infty} \left(\sum_{j=0}^{T-1} a_{k0j}^{\nu} \exp\left(2\pi ij(n-k)/T\right) \right) X_{n-k}, \tag{6}$$

where

$$\underline{A}_k^{\nu} = (a_{klj}^{\nu})_{l,j=0}^{T-1}.$$

It is interesting to note that the coefficients a_{k0j}^{ν} in (6) are exactly those in the Fourier decomposition of the coefficients in the periodically varying autoregressive prediction formula for the PCS X_n.

Before proving Theorem 4.1 we need to prove two more lemmas.

LEMMA 4.2. *For each n, the subspace $H(\underline{Z}; n)$ of $K = H \oplus \cdots \oplus H$ is of the form*

$$H(\underline{Z}; n) = H(X; n) \oplus H(X; n-1) \oplus \cdots \oplus H(X; n-T+1).$$

Proof. Take any $0 \leqq k \leqq T-1$ and $m \leqq n$ and consider the linear combination

$$\sum_{j=0}^{T-1} \sqrt{T}\, Z_m^j \exp\left(-2\pi ij(m+k)/T\right)$$

in $H(\underline{Z}; n)$. We can write it as

$$\begin{aligned}
&\sum_{j=0}^{T-1} \sqrt{T}\, Z_m^j \exp\left(-2\pi ij(m+k)/T\right) \\
&\quad = \sum_{j=0}^{T-1} \left(\bigoplus_{l=0}^{T-1} X_{m+l} \exp\left(2\pi ij(m+l)/T\right) \exp\left(-2\pi ij(m+k)/T\right) \right) \\
&\quad = \bigoplus_{l=0}^{T-1} X_{m+l} \left(\sum_{j=0}^{T-1} \exp\left(2\pi ij(l-k)\right) \right) \\
&\quad = \bigoplus_{l=0}^{T-1} T X_{m+l} \delta_{lk}.
\end{aligned}$$

This proves that

$$0\oplus\cdots\oplus\underset{\substack{\uparrow\\ j\text{th term}}}{H(X;\,n+j)}\oplus 0\cdots\oplus 0\subseteq H(\underline{Z};\,n)\quad\text{for each } 0\leqq j\leqq T-1.$$

This in turn shows that

$$H(X;\,n)\oplus H(X;\,n-1)\oplus\cdots\oplus H(X;\,n+T-1)\subseteq H(\underline{Z};\,n)\quad\text{for each } n.$$

The other inclusion is an immediate consequence of the fact that for each $0\leqq j\leqq T-1$ and each $m\leqq n$,

$$\sqrt{T}\,Z_m^j=\bigoplus_{k=0}^{T-1}X_{m+k}\exp\left(2\pi ij(m+k)/T\right)$$

belongs to $H(X;\,n)\oplus H(X;\,n+1)\oplus\cdots\oplus H(X;\,n+T-1)$. □

LEMMA 4.3. *For each n, the subspace $\underline{H}(\underline{Z};\,n)$ of K^T can be expressed as $\underline{H}(\underline{Z};\,n)=(H(\underline{Z};\,n))^T$.*

Proof. Take any finite linear combination

$$\underline{Z}=\sum_{k\leqq n}\underline{A}_k\underline{Z}_k$$

in $\underline{H}(\underline{Z};\,n)$, and consider its ith component Z^i, $0\leqq i\leqq T-1$. We have

$$Z^i=\sum_{k\leqq n}\sum_{j=0}^{T-1}(A_k)_{i,j}Z_k^j.$$

Clearly $Z^i\in H(\underline{Z};\,n)$, which implies $\underline{Z}\in(H(\underline{Z};\,n))^T$. Thus

$$sp\{\underline{Z}_k:\,k\leqq n\}\subseteq(H(\underline{Z};\,n))^T.$$

Taking closure on the left-hand side we get $\underline{H}(\underline{Z};\,n)\subseteq(H(\underline{Z};\,n))^T$. The other inclusion follows from a similar argument.

Proof of the Theorem 4.1. Considering the 0th component $Z_{n,\nu}^0$ of the predictor $\underline{Z}_{n,\nu}$, we can write

$$\begin{aligned}Z_{n,\nu}^0&=(\underline{Z}_{n+\nu}\,|\,\underline{H}(\underline{Z};\,n))^0\\&=(Z_{n,\nu}^0\,|\,H(\underline{Z};\,n))\\&=\left(\bigoplus_{s=0}^{T-1}X_{n+\nu+s}\,\middle|\,\bigoplus_{s=0}^{T-1}H(X;\,n+s)\right)\\&=\bigoplus_{s=0}^{T-1}(X_{n+\nu+s}\,|\,H(X;\,n+s)).\end{aligned}$$

(The second equality follows from Lemma 4.3 and our earlier comment on the relation between projection in K^T and the projection in K. The third equality follows from Lemma 4.2.)

On the other hand we can write

$$\begin{aligned}Z_{n,\nu}^0&=\left(\sum_{k=0}^{\infty}\underline{A}_K^\nu\underline{Z}_{n-k}\right)^0=\sum_{k=0}^{\infty}\left(\sum_{j=0}^{T-1}a_{k0j}^\nu Z_{n-k}^j\right)\\&=\frac{1}{\sqrt{T}}\sum_{k=0}^{\infty}\sum_{j=0}^{T-1}a_{k0j}^\nu\left(\bigoplus_{s=0}^{T-1}X_{n-k+s}\exp\left(2\pi ij(n-k)/T\right)\right)\\&=\bigoplus_{s=0}^{T-1}\frac{1}{\sqrt{T}}\left[\sum_{k=0}^{\infty}\left(\sum_{j=0}^{T-1}a_{k0j}^\nu\exp\left(2\pi ij(n-k)/T\right)\right)X_{n-k+s}\right].\end{aligned}$$

Hence by comparison we get

$$(X_{n+\nu}|H(X;n))=\sum_{k=0}^{\infty}\sum_{j=0}^{T-1}\frac{1}{\sqrt{T}}a_{k0j}^{\nu}\exp(2\pi ij(n-k)/T)X_{n-k},$$

which completes the proof. □

Now we will find a formula for the prediction error $g^{\nu}(n)=\|X_{n+\nu}-X_{n,\nu}\|$ of a PCS X_n. In contrast to the stationary case where the prediction error is independent of time n, here the prediction error $g^{\nu}(n)$ does vary with time. However, it varies in a periodic way. So, $g^{\nu}(n)$ can be written as

$$g^{\nu}(n)=\sum_{k=0}^{T-1}g_k^{\nu}\exp(2\pi ikn/T).$$

By inverting this we get

$$g_k^{\nu}=\frac{1}{T}\sum_{n=0}^{T-1}g^{\nu}(n)\exp(-2\pi ink/T).$$

Using this we can write

$$\begin{aligned}
Tg_k^{\nu}&=\sum_{n=0}^{T-1}(X_{n+\nu}-X_{n,\nu},X_{n+\nu}-X_{n,\nu})\exp(-2\pi ink/T)\\
&=\sum_{n=0}^{T-1}(X_{n+\nu}-X_{n,\nu},(X_{n+\nu}-X_{n,\nu})\exp(2\pi ink/T))\\
&=\sum_{n=0}^{T-1}(X_{n+\nu}-X_{n,\nu},(X_{n+\nu}-X_{n,\nu})\exp(2\pi ik(n+\nu)/T))\exp(2\pi ik\nu/T)\\
&=T\exp(2\pi ik\nu/T)\langle Z_{\nu}^{0}-Z_{0,\nu}^{0},Z_{\nu}^{k}-Z_{0,\nu}^{k}\rangle.
\end{aligned}$$

Dividing through by T we get

$$g_k^{\nu}=\exp(2\pi ik\nu/T)G_{0k}^{\nu}.$$

We have thus proven the following theorem, which provides a formula for the prediction error of our PSC X_n.

THEOREM 4.4. *With the notation of Theorem* 4.1 *the prediction error* $g^{\nu}(n)$ *of a* PCS X_n *can be written as*

$$g^{\nu}(n)=\sum_{k=0}^{T-1}G_{0k}^{\nu}\exp(2\pi ik(n+\nu)/T), \tag{7}$$

where

$$\underline{G}^{\nu}=(G_{kk'}^{\nu})_{k,k'=0}^{T-1}$$

is the ν*-step prediction error matrix of predicting its associate* SS $\underline{Z}_n$ *in* K^T.

In the rest of this section we present an algorithm for determining the best linear predictor and the prediction error of a PCS X_n under the following assumption.

Assumption 4.5. Suppose X_n is a PCS whose corresponding spectral measure $d\underline{F}_X$ given by (4) is absolutely continuous with respect to the Lebesgue measure and

(i) Its derivative $\underline{F}'_X(\lambda)$ is almost everywhere invertible;

(ii) All entries of $\underline{F}'_X(\lambda)$ belong to L^{∞};

(iii) All entries of $\underline{F}'_X(\lambda)^{-1}$ belong to L_1.

PROPOSITION 4.6. *If the* PCS X_n *satisfies the requirement of Assumption* 4.5, *then its associate* SS $\underline{Z}_n$ *in* K^T *given in Theorems* 1.1 *and* 3.1, *considered as a multivariate process with components in the Hilbert space K, satisfies Masani's conditions in* § 5 *of* [5].

Proof. By Lemma 3.3 the matricial spectral measure $\underline{F}_z$ of $\underline{Z}_n$ is exactly the same as the spectral measure $\underline{F}_X$. On the other hand, Masani's conditions in § 5 of [5] are exactly those given in Assumption 4.5. This completes the proof. □

Thus under Assumption 4.5 the matricial density function $\underline{F}'_X$ admits a factorization of the form

$$\underline{F}'_X(\lambda) = \underline{\Phi}(\lambda)\underline{\Phi}^*(\lambda),$$

where the optimal factor $\underline{\Phi}$ and its inverse $\underline{\Phi}^{-1}$ have square integrable entries whose negative Fourier coefficients are zero, i.e.,

$$\underline{\Phi}(\lambda) = \sum_{k=0}^{\infty} \underline{C}_k \exp(ik\theta), \qquad \underline{\Phi}^{-1}(\theta) = \sum_{k=0}^{\infty} \underline{D}_k \exp(ik\theta).$$

(For this and other important results about prediction of multivariate SS's see [5], [6].) It is also well known (cf. [5], [6]) that the corresponding T-variate SS $\underline{Z}_n$ has an autoregressive representation of the form (5) with $\underline{A}_k^\nu$ given by

$$\underline{A}_k^\nu = \sum_{n=0}^{k} \underline{C}_{\nu+n}\underline{D}_{k-n}. \tag{8}$$

Under these conditions there is actually an algorithm developed in [6] which allows us to find the matrix coefficients $\underline{C}_k$'s and $\underline{D}_k$'s from the Fourier coefficients of the matricial function $\underline{M}(\lambda) = \underline{F}'_X(\lambda) - \underline{I}$. Hence, we can first find $\underline{C}_k$'s and $\underline{D}_k$'s and put them in (8) to find the autoregressive coefficients $\underline{A}_k^\nu$. Then we can use relation (5) to find the best linear predictor $X_{n,\nu}$ of the given PCS X_n. In a similar fashion we can use the results of [6] to find the prediction error matrix $\underline{G}^\nu$ of the ν-step prediction of the multivariate SS $\underline{Z}_n$ (which is given in terms of $\underline{C}_k$'s and $\underline{D}_k$'s) and use those in conjunction with (7) to find the prediction errors of the original PCS X_n.

Acknowledgment. The author is glad to acknowledge the referees' comments which improved the original version of this paper.

REFERENCES

[1] E. G. Gladyshev, *Periodically correlated random sequences*, Soviet Math. Dokl., 2 (1961), pp. 385-388.

[2] L. I. Gudzenko, *On periodically nonstationary processes*, Radiotekhn, i Elektron, 6 (1959), pp. 1062-1064.

[3] H. L. Hurd, *Periodically correlated processes with discontinuous correlation functions*, Theory Probab. Appl., 19 (1974), pp. 834-838.

[4] ———, *Representation of harmonizable periodically correlated processes and their covariances*, J. Multivariate Anal., 29 (1989), pp. 53-67.

[5] P. Masani, *The prediction theory of multivariate stochastic processes* III, Acta Math., 104 (1960), pp. 141-162.

[6] ———, *Recent trends in multivariate prediction theory*, in Multivariate Analysis, Proceedings of an International Symposium, P. R. Krishnaiah, ed., Academic Press, New York, 1966, pp. 351-382.

[7] A. G. Miamee, *Periodically correlated processes and their stationary dilations*, SIAM J. Appl. Math., 50 (1990), pp. 1194-1199.

[8] A. G. Miamee and H. Salehi, *On the prediction of periodically correlated stochastic processes*, In Multivariate Anal. V, P. R. Krishnaiah, ed., North-Holland, Amsterdam, 1980, pp. 167-179.

[9] N. Wiener and P. Masani, *The prediction theory of multivariate stochastic processes* I, Acta. Math., 98 (1957), pp. 111-150.

[10] ———, *The prediction theory of multivariate stochastic processes* II, Acta Math., 99 (1958), pp. 93-137.

SIAM J. MATH. ANAL.
Vol. 24, No. 3, pp. 712–738, May 1993

011

NONORTHOGONAL WAVELET PACKETS*

CHARLES K. CHUI† AND CHUN LI‡

Abstract. The notion of orthonormal wavelet packets introduced by Coifman and Meyer is generalized to the nonorthogonal setting in order to include compactly supported and symmetric basis functions. In particular, dual (or biorthogonal) wavelet packets are investigated and a stability result is established. Algorithms for implementations are also developed.

Key words. multiresolution analysis, scaling functions, wavelets, dual wavelets, wavelet packets, frames, tree algorithms, decomposition and reconstruction algorithms

AMS subject classifications. primary 41A58; secondary 42C30

1. Introduction.

Orthogonal wavelet packets (also called wave packets) introduced by Coifman and Meyer [5] (see also [6], [11]) are used to further decompose wavelet components. Based on Daubechies' compactly supported orthogonal wavelets [7], procedures for both computations and implementation can be made very efficient. However, the intrinsic property of lack of symmetry persists. This causes phase distortion in applications that require lossy data compression and decompression. More recently, compactly supported symmetric wavelets are available by sacrificing orthogonality partially [2] or totally [4]. A unified treatment of such wavelets is given in [1, Thm. 5.19]. Thus, it is natural to extend the study of orthonormal wavelet packets to the nonorthogonal setting. This extension is valuable because linear-phase filters (resulting from symmetric wavelets) cannot be constructed by using compactly supported orthogonal wavelets, but can be constructed by using semi-orthogonal or biorthogonal ones. In addition, wavelet packets provide better frequency localization than wavelets while time-domain localization is not lost [6], [11]. This capability enhances the application of wavelet decomposition in processing signals with high-frequency components. Of course when nonorthogonal wavelets are being considered, two very fundamental problems must be considered, namely, the existence of duals and the stability requirement. The difficulty of these problems propagates when the notion of orthogonal wavelet packets is generalized to the nonorthogonal setting. The objective of this paper is to give a careful treatment of this generalization, with special emphasis on the consideration of dual wavelet packets, stability, and development of algorithms. We emphasize that our starting point is based on the work of Coifman and Meyer [5].

Throughout this paper, the space of all square-integrable functions on the real line will be denoted, as usual, by $L^2 := L^2(\mathbb{R})$, and the notation for inner product and Fourier transform of functions in L^2 is given by

$$\langle f, g\rangle = \int_{-\infty}^{\infty} f(x)\overline{g(x)}\,dx \tag{1.1}$$

*Received by the editors January 28, 1992; accepted for publication (in revised form) August 5, 1992. This research was supported by National Science Foundation grant DMS 89-01345 and Army Research Office contract DAAL 03-90-G-0091.

†Department of Mathematics, Texas A&M University, College Station, Texas 77843.

‡Institute of Mathematics, Academia Sinica, Beijing, 100080, People's Republic of China. Present address, Department of Mathematics, Texas A&M University, College Station, Texas 77843.

and

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} e^{-i\omega x} f(x)\, dx, \tag{1.2}$$

respectively. Also the norm of any f in L^2 will be denoted by $\|f\| = \langle f, f\rangle^{1/2}$; and for any function f, we will always use the notation

$$f_{j,k}(x) := 2^{j/2} f(2^j x - k). \tag{1.3}$$

A function $\psi \in L^2$ will be called a "wavelet" if there exists another function $\widetilde{\psi} \in L^2$, called the "dual wavelet" of ψ, such that the family

$$\{\psi_{j,k}\colon\ j,k \in \mathbb{Z}\}$$

is a Riesz (or unconditional) basis of L^2, and the collection

$$\{\widetilde{\psi}_{j,k}\colon\ j,k \in \mathbb{Z}\}$$

is the corresponding dual (or biorthogonal) basis in the sense that

$$\langle \psi_{j,k}, \widetilde{\psi}_{j',k'}\rangle = \delta_{j,j'}\delta_{k,k'} \tag{1.4}$$

for all $j, k, j', k' \in \mathbb{Z}$. In this paper, when we call a family a "basis" of any subspace of L^2, we always mean that it is a Schauder basis of this subspace. Unless we specify explicitly, these Schauder bases need not be Riesz bases. Recall that a Riesz basis is a Schauder basis satisfying the additional condition that the L^2-norm of any series representation and the ℓ^2-norm of its coefficient sequence are equivalent. We shall call this extra requirement the "stability" condition. A very powerful tool for constructing wavelets is the consideration of multiresolution analyses (MRA) of L^2, introduced by Meyer [10] and Mallat [9]. A function that "generates" any MRA of L^2 is called a "scaling function."

In the study of scaling functions and wavelets, the symbol of a sequence is often used. For convenience, we will usually consider symbols

$$A(z) = \sum_{n \in \mathbf{Z}} a_n z^n$$

of sequences $\{a_n\}$ in ℓ^1, namely, $\Sigma |a_n| < \infty$. The collection of such Laurent series is called the Wiener class, which will be denoted by $\mathcal{W}$.

The outline of this paper is as follows. In §2, we will introduce the necessary notations and definitions and state the main results of this paper. Of particular importance is that if $\psi_1 := \psi$ is a wavelet with dual $\widetilde{\psi}_1 := \widetilde{\psi}$, so that both $\{\psi_{j,k}\}$ and $\{\widetilde{\psi}_{j,k}\}$ are Riesz bases of L^2, then for each nonnegative integer ℓ, both

$$\{\psi_{n;j,k}\colon\ j,k \in \mathbb{Z},\ 2^\ell \le n < 2^{\ell+1}\} \tag{1.5}$$

and its dual family

$$\{\widetilde{\psi}_{n;j,k}\colon\ j,k \in \mathbb{Z},\ 2^\ell \le n < 2^{\ell+1}\} \tag{1.6}$$

are also Reisz bases of L^2. Here, $\{\psi_n\}$ is the sequence of "wavelet packets" induced by the wavelet ψ and its corresponding scaling function $\psi_0 := \phi$, and $\{\tilde{\psi}_n\}$ denotes the corresponding sequence of dual wavelet packets. This is a consequence of Theorems 3 and 6. It is interesting to point out that only finitely many ψ_n's are used in the family (1.5), and that when infinitely many ψ_n's are considered, then stability (or unconditionality) may be lost in general. This negative result is due to Cohen and Daubechies [3] and was brought to our attention by Albert Cohen. Preliminary results in this paper, which are perhaps of some independent interest, are derived in §3, and proofs of the main results are given in §4 . In §5, two algorithms for implementation will be developed.

2. Main results. If P^0 and G^0 are Laurent series in the Wiener class $\mathcal{W}$ satisfying

$$\begin{aligned} P^0(z) &= \left(\frac{1+z}{2}\right)^N S(z), \\ G^0(z) &= \left(\frac{1+z}{2}\right)^{\tilde{N}} \tilde{S}(z), \end{aligned} \tag{2.1}$$

$$S(1) = \tilde{S}(1) = 1, \tag{2.2}$$

and

$$\begin{aligned} &\inf_{j>0} \max_{\omega} \prod_{k=1}^{j} |S(e^{i2^{-k}\omega})|^{\frac{1}{j}} < 2^{N-\frac{1}{2}}, \\ &\inf_{j>0} \max_{\omega} \prod_{k=1}^{j} |\tilde{S}(e^{i2^{-k}\omega})|^{\frac{1}{j}} < 2^{\tilde{N}-\frac{1}{2}}, \end{aligned} \tag{2.3}$$

where N and $\tilde{N}$ are positive integers, then the infinite products

$$\begin{aligned} \hat{\phi}(\omega) &:= \prod_{k=1}^{\infty} P^0(e^{-i\omega/2^k}); \\ \hat{\tilde{\phi}}(\omega) &:= \prod_{k=1}^{\infty} \overline{G^0(e^{-i\omega/2^k})} \end{aligned} \tag{2.4}$$

converge in L^2, and the limit functions are Fourier transforms of some functions ϕ, $\tilde{\phi} \in L^2$, that generate two (possibly different) multiresolution analyses (MRA) $\{V_n\}$ and $\{\tilde{V}_n\}$, respectively, of L^2 (see [4] and [1, Thm. 5.22]). In addition, from (2.4), it follows that $\hat{\phi}$ and $\hat{\tilde{\phi}}$ satisfy

$$\begin{aligned} \hat{\phi}(\omega) &= P^0(z)\hat{\phi}\left(\frac{\omega}{2}\right), \\ \hat{\tilde{\phi}}(\omega) &= \overline{G^0(z)}\,\hat{\tilde{\phi}}\left(\frac{\omega}{2}\right), \end{aligned} \tag{2.5}$$

where $z = e^{-i\omega/2}$; and writing

$$\begin{aligned} P^0(z) &= \frac{1}{2}\sum_{n\in\mathbf{Z}} p_n^0 z^n, \\ G^0(z) &= \frac{1}{2}\sum_{n\in\mathbf{Z}} g_n^0 z^n, \end{aligned} \tag{2.6}$$

we see that (2.5) is equivalent to the "two-scale relations"

$$\begin{aligned}\phi(x) &= \sum_{k\in\mathbf{Z}} p_k^0 \phi(2x-k),\\ \tilde{\phi}(x) &= \sum_{k\in\mathbf{Z}} \overline{g_{-k}^0}\tilde{\phi}(2x-k),\end{aligned} \tag{2.7}$$

of the "scaling functions" ϕ and $\tilde{\phi}$. Hence, the sequences $\{p_k^0\}$ and $\{\overline{g_{-k}^0}\}$ are the corresponding "two-scale sequences," and $P^0(z)$ and $\overline{G^0(z)}$ the corresponding "two-scale symbols," of the scaling functions ϕ and $\tilde{\phi}$, respectively.

We will say that ϕ and $\tilde{\phi}$ are dual to each other if they satisfy

$$\langle \phi(\cdot - j), \tilde{\phi}(\cdot - k)\rangle = \delta_{j,k}, \qquad j,k \in \mathbb{Z}. \tag{2.8}$$

Under the assumptions (2.1)–(2.3), a necessary and sufficient condition for the duality relationship (2.8) is that P^0 and G^0 are "dual two-scale symbols," in the sense that

$$P^0(z)G^0(z) + P^0(-z)G^0(-z) = 1, \qquad |z| = 1. \tag{2.9}$$

A proof of this statement is given in [1, Thm. 5.22] (see also [4] for the case of polynomial symbols). It should be noted that the assumptions in (2.1)–(2.3) for P^0 and G^0 can be somewhat weakened. Hence, for more generality, we will drop this requirement, but simply assume that $\hat{\phi}$ and $\hat{\tilde{\phi}}$ defined in (2.4) are in L^2, $\{V_n\}$ and $\{\tilde{V}_n\}$ are MRA of L^2, and that (2.8) and (2.9) are satisfied.

Next, let us consider an arbitrary Laurent series $R(z)$ of class $\mathcal{W}$, which never vanishes on the unit circle $|z| = 1$. By Wiener's lemma, we also have $1/R(z) \in \mathcal{W}$, and this yields two other Laurent series in $\mathcal{W}$, namely,

$$\begin{aligned}P^1(z) &:= -zG^0(-z)R(z^2),\\ G^1(z) &:= -z^{-1}P^0(-z)/R(z^2).\end{aligned} \tag{2.10}$$

The reason for introducing P^1 and G^1 is that the two matrices

$$\begin{aligned}M(z) &:= \begin{bmatrix} P^0(z) & P^1(z)\\ P^0(-z) & P^1(-z)\end{bmatrix},\\ \widetilde{M}(z) &:= \begin{bmatrix} G^0(z) & G^1(z)\\ G^0(-z) & G^1(-z)\end{bmatrix}\end{aligned} \tag{2.11}$$

are nonsingular for $|z| = 1$, independent of the choice of R, and that $M^T(z)$ and $\widetilde{M}(z)$ are inverses of each other. A discussion of this result will be given in the next section. Now, we write

$$\begin{aligned}P^1(z) &= \frac{1}{2}\sum_{n\in\mathbf{Z}} p_n^1 z^n,\\ G^1(z) &= \frac{1}{2}\sum_{n\in\mathbf{Z}} g_n^1 z^n,\end{aligned} \tag{2.12}$$

and set

$$\psi_0 := \phi \quad \text{and} \quad \tilde{\psi}_0 := \tilde{\phi}. \tag{2.13}$$

Then, in view of (2.7), following Coifman and Meyer [5] we may introduce two sequences of L^2-functions, $\{\psi_n\}$ and $\{\widetilde{\psi}_n\}$, defined by

$$(2.14)\qquad \begin{aligned} \psi_{2n+\lambda}(x) &:= \sum_{k\in\mathbf{Z}} p_k^\lambda \psi_n(2x-k), \quad \lambda = 0,1, \\ \widetilde{\psi}_{2n+\lambda}(x) &:= \sum_{k\in\mathbf{Z}} \overline{g_{-k}^\lambda}\, \widetilde{\psi}_n(2x-k), \quad \lambda = 0,1, \end{aligned}$$

where $n = 0, 1, \dots$.

Of course, (2.14) reduces to (2.7) when $\lambda = 0$ and $n = 0$. The functions

$$(2.15)\qquad \psi := \psi_1 \quad \text{and} \quad \widetilde{\psi} := \widetilde{\psi}_1,$$

obtained by setting $\lambda = 1$ and $n = 0$ in (2.14), are dual wavelets in the sense of (1.4); that is,

$$(2.16)\qquad \langle \psi_{j,k}, \widetilde{\psi}_{\ell,m}\rangle = \langle \psi_{1;j,k}, \widetilde{\psi}_{1;\ell,m}\rangle = \delta_{j,\ell}\delta_{k,m}, \quad j,k,\ell,m \in \mathbb{Z}.$$

We will document this statement in a moment. In general, we call $\{\psi_n\}$ and $\{\widetilde{\psi}_n\}$ sequences of *wavelet packets*, and $\{\widetilde{\psi}_n\}$ the dual of $\{\psi_n\}$. In the orthonormal setting, where $\widetilde{\psi}_n = \psi_n$, $n = 0, 1, \dots$ (obtained by considering $\overline{G^\lambda(z)} = P^\lambda(z)$ and $|R(z)| = 1$ for $|z| = 1$ and $\lambda = 0, 1$, in (2.10)), the functions ψ_n become the orthonormal wavelet packets introduced by Coifman and Meyer [5]. (See also [6], [11], and [1, Chap. 7].) The duality of the scaling functions in (2.8) and that of the wavelets ψ and $\widetilde{\psi}$ (as defined by (2.15)), for $j = \ell = 0$ in (2.16), are generalized to wavelet packets in this paper as follows.

THEOREM 1. *Assume that $\hat{\phi}$ and $\hat{\tilde{\phi}}$ defined in* (2.4) *are in L^2 and generate two* MRA *of L^2, and that* (2.8) *and* (2.9) *are satisfied. Then, for all $m, n = 0, 1, \dots,$ and $k, \ell \in \mathbb{Z}$,*

$$(2.17)\qquad \langle \psi_m(\cdot - k), \widetilde{\psi}_n(\cdot - \ell)\rangle = \delta_{m,n}\delta_{k,\ell}.$$

Let us turn to a discussion of the MRA generated by the scaling functions $\psi_0 = \phi$ and $\widetilde{\psi}_0 = \widetilde{\phi}$, namely,

$$(2.18)\qquad \begin{cases} V_j := \operatorname{clos}_{L^2}\langle \psi_{0;j,k}\colon\ k \in \mathbb{Z}\rangle, \\ \widetilde{V}_j := \operatorname{clos}_{L^2}\langle \widetilde{\psi}_{0;j,k}\colon\ k \in \mathbb{Z}\rangle, \end{cases}$$

and the (complementary) wavelet spaces

$$(2.19)\qquad \begin{cases} W_j := \operatorname{clos}_{L^2}\langle \psi_{1;j,k}\colon\ k \in \mathbb{Z}\rangle, \\ \widetilde{W}_j := \operatorname{clos}_{L^2}\langle \widetilde{\psi}_{1;j,k}\colon\ k \in \mathbb{Z}\rangle, \end{cases}$$

generated by the wavelets $\psi = \psi_1$ and $\widetilde{\psi} = \widetilde{\psi}_1$. Recall that

$$(2.20)\qquad L^2 = \operatorname{clos}_{L^2}\left(\bigcup_{j\in\mathbf{Z}} V_j\right) = \operatorname{clos}_{L^2}\left(\bigcup_{j\in\mathbf{Z}} \widetilde{V}_j\right),$$

and that for each $j \in \mathbb{Z}$,

$$\begin{cases} V_{j+1} = V_j \dot{+} W_j, \\ \widetilde{V}_{j+1} = \widetilde{V}_j \dot{+} \widetilde{W}_j, \end{cases} \tag{2.21}$$

where $\dot{+}$ denotes direct-sum decomposition, and the duality condition is reflected by the orthogonality property

$$V_j \perp \widetilde{W}_j \quad \text{and} \quad \widetilde{V}_j \perp W_j, \qquad j \in \mathbb{Z}. \tag{2.22}$$

(See [1, Chap. 5] for more details.) As a consequence of (2.20) and (2.21), and the fact that

$$\bigcap_{j\in\mathbb{Z}} V_j = \bigcap_{j\in\mathbb{Z}} \widetilde{V}_j = \{0\},$$

we have

$$L^2 = \dot{\sum_{j\in\mathbb{Z}}} W_j = \dot{\sum_{j\in\mathbb{Z}}} \widetilde{W}_j. \tag{2.23}$$

Here and throughout, the "infinite direct-sum" $\dot{\sum}$ is defined, as usual, to be the L^2-closure of the truncated direct-sums:

$$\dot{\sum_{-m\le j\le n}}, \qquad m,\ n \in \mathbb{Z}_+.$$

More precisely, we have

$$\dot{\sum_{j\in\mathbb{Z}}} W_j = \operatorname{clos}_{L^2}\left(\bigcup_{m,n\in\mathbb{Z}_+} \dot{\sum_{-m\le j\le n}} W_j \right).$$

We will elaborate on the existence and uniqueness of the infinite direct-sum decompositions of functions $f \in L^2$ at the end of the next section. To further decompose the wavelet spaces W_j and $\widetilde{W}_j$, we consider the linear spaces

$$\begin{cases} U_n := \operatorname{clos}_{L^2}\langle \psi_n(\cdot - k)\colon\ k \in \mathbb{Z}\rangle, \\ \widetilde{U}_n := \operatorname{clos}_{L^2}\langle \widetilde{\psi}_n(\cdot - k)\colon\ k \in \mathbb{Z}\rangle, \end{cases} \tag{2.24}$$

and obtain the following result.

THEOREM 2. *For each $j = 0, 1, \ldots,$*

$$V_j = \dot{\sum_{0\le n<2^j}} U_n, \qquad W_j = \dot{\sum_{2^j\le n<2^{j+1}}} U_n, \tag{2.25}$$

$$\widetilde{V}_j = \dot{\sum_{0\le n<2^j}} \widetilde{U}_n, \qquad \widetilde{W}_j = \dot{\sum_{2^j\le n<2^{j+1}}} \widetilde{U}_n. \tag{2.26}$$

Consequently,

$$L^2 = U_0 \dot{+} \sum_{j\in\mathbb{Z}_+}^{\bullet} \sum_{2^j \le n < 2^{j+1}}^{\bullet} U_n = \widetilde{U}_0 \dot{+} \sum_{j\in\mathbb{Z}_+}^{\bullet} \sum_{2^j \le n < 2^{j+1}}^{\bullet} \widetilde{U}_n, \tag{2.27}$$

where, as in (2.23), *the infinite direct-sums are L^2-closures of the limits of the corresponding one-sided truncated direct-sums. Moreover,*

$$U_m \perp \widetilde{U}_n, \quad m \neq n, \quad m,n \in \mathbb{Z}_+, \tag{2.28}$$

and $\{\psi_m(\cdot - k)\colon\ k \in \mathbb{Z}\}$, $\{\widetilde{\psi}_n(\cdot - k)\colon\ k \in \mathbb{Z}\}$ are (biorthogonal) bases of U_m, $\widetilde{U}_n$, respectively, where $m,n \in \mathbb{Z}_+$.

It should be noted that the direct-sum decompositions in (2.27) cannot be written as $L^2 = \sum^{\bullet}_{n\in\mathbb{Z}_+} U_n$, in general, in the sense that the truncated projection operators may not converge strongly. The reason is that the families $\{\psi_m(\cdot - k)\colon\ k \in \mathbb{Z},\ m \in \mathbb{Z}_+\}$ and $\{\widetilde{\psi}_n(\cdot - k)\colon\ k \in \mathbb{Z},\ n \in \mathbb{Z}_+\}$ may not be Riesz bases. Nevertheless, the sequence of projection operators corresponding to truncation of the partial sum at $n = 2^j - 1$ converges strongly. This will be discussed at the end of next section. The biorthogonality between the wavelet $\psi = \psi_1$ and its dual $\widetilde{\psi} = \widetilde{\psi}_1$ as described by (2.16) can be generalized to wavelet packets as follows.

THEOREM 3. *Let $\ell \in \mathbb{Z}_+$ be arbitrarily chosen. Then both of the families*

$$\{\psi_{n;j,k}\colon\ j,k \in \mathbb{Z}, \quad 2^\ell \le n < 2^{\ell+1}\} \tag{2.29}$$

and

$$\{\widetilde{\psi}_{n;j,k}\colon\ j,k \in \mathbb{Z}, \quad 2^\ell \le n < 2^{\ell+1}\} \tag{2.30}$$

are bases of L^2, and they are biorthogonal in the sense that

$$\langle \psi_{n;j,k}, \widetilde{\psi}_{n';j',k'} \rangle = \delta_{n,n'}\delta_{j,j'}\delta_{k,k'} \tag{2.31}$$

for all $j,k,j',k',n,n' \in \mathbb{Z}$ with $2^\ell \le n,n' < 2^{\ell+1}$.

Of course, assertion (2.31) reduces to (2.16) when $\ell = 0$. In the next theorem, we will allow the (scaling) index j in (2.29) and (2.30) to depend on the index n, and obtain another pair of biorthogonal bases of L^2. For this purpose, we introduce the notation

$$\ell(n) := \lfloor \log_2 n \rfloor, \tag{2.32}$$

where $\lfloor x \rfloor$ denotes, as usual, the largest integer not exceeding x.

THEOREM 4. *Both of the families*

$$\{\psi_{0;0,k}, \psi_{n;\ell(n),k}, \psi_{n;\ell(n)+1,k}\colon\ k,n \in \mathbb{Z}, n \ge 1\} \tag{2.33}$$

and

$$\{\widetilde{\psi}_{0;0,k}, \widetilde{\psi}_{n;\ell(n),k}, \widetilde{\psi}_{n;\ell(n)+1,k}\colon\ k,n \in \mathbb{Z}, n \ge 1\} \tag{2.34}$$

are bases of L^2, and they are biorthogonal in the sense of (2.31).

There is, however, a common theme among the biorthogonal bases in Theorems 2–4, and a unified result can be stated as follows.

THEOREM 5. *Let $\mathcal{J}$ be a collection of ordered pairs (n, j), where $n \in \mathbb{Z}_+$, $j \in \mathbb{Z}$, such that the dyadic intervals*

$$I_{n,j} := [2^j n, 2^j(n+1)) \tag{2.35}$$

form a disjoint covering of the interval $(0, \infty)$, and that every bounded subinterval of $(0, \infty)$ is contained in the union of finitely many $I_{n,j}$. Then both of the families

$$\{\psi_{n;j,k}\colon\ (n,j) \in \mathcal{J}, \quad k \in \mathbb{Z}\} \tag{2.36}$$

and

$$\{\tilde{\psi}_{n;j,k}\colon\ (n,j) \in \mathcal{J}, \quad k \in \mathbb{Z}\} \tag{2.37}$$

are bases of L^2, and they are biorthogonal in the sense of (2.31) *for all (n, j), (n', j') in $\mathcal{J}$ and $k, k' \in \mathbb{Z}$.*

In the following, we will discuss the "stability" of the biorthogonal bases in (2.29) and (2.30). Since we already have the basis structure, it is sufficient to show that $\{\psi_{n;j,k}\}$ and $\{\tilde{\psi}_{n;j,k}\}$ in (2.29) and (2.30) are "dual frames," where duality is guaranteed by (2.31). Recall that a family $\{f_j\colon\ j \in J\}$ is said to constitute a frame of L^2 if there exist positive constants A and B such that

$$A\|f\|^2 \le \sum_{j\in J} |\langle f, f_j\rangle|^2 \le B\|f\|^2, \qquad f \in L^2 \tag{2.38}$$

(cf. [8], [7], and [1, Chap. 3]).

THEOREM 6. *Suppose that both*

$$\{\psi_{1;j,k}\colon\ j,k \in \mathbb{Z}\} \quad \textit{and} \quad \{\tilde{\psi}_{1;j,k}\colon\ j,k \in \mathbb{Z}\} \tag{2.39}$$

are frames of L^2. Then for any $\ell \in \mathbb{Z}_+$, both of the families in (2.29) *and* (2.30) *are also frames of L^2.*

In other words, if the families in (2.29) and (2.30) with $\ell = 0$ are frames of L^2, then they are frames of L^2 for any positive integer ℓ. However, the new frame bounds are usually exponential in ℓ as can be seen later from the proof of Theorem 6. Recall from Theorem 3 that each of the families (2.29) and (2.30) is "linearly independent" in the sense that no $\psi_{n';j',k'}$, $2^\ell \le n' < 2^{\ell+1}$ and $j', k' \in \mathbb{Z}$, lies in the L^2-closure of the (finite) linear span of

$$\psi_{n;j,k}, \qquad (n,j,k) \ne (n',j',k'),$$

where $2^\ell \le n < 2^{\ell+1}$ and $j, k \in \mathbb{Z}$. Hence, by a result on frames in [8] (see also [4]), it follows from Theorem 6 that both of the families in (1.5) and (1.6) (or (2.29) and (2.30)) are Riesz bases of L^2, provided, of course, that ψ is a wavelet with dual $\tilde{\psi}$.

To establish the results stated in this section, we need a sequence of lemmas that are of some independent interest.

3. Auxiliary results.

We first observe that the Laurent series P^λ and G^λ, $\lambda = 0, 1$, that satisfy (2.9) and (2.10) for some $R \in \mathcal{W}$ with $R(z) \ne 0$ on $|z| = 1$, also satisfy

$$\begin{aligned} &P^0(z)G^0(z) + P^1(z)G^1(z) = 1, \\ &P^0(-z)G^0(z) + P^1(-z)G^1(z) = 0, \qquad |z| = 1, \end{aligned} \tag{3.1}$$

and

$$P^\lambda(z)G^\mu(z)+P^\lambda(-z)G^\mu(-z)=\delta_{\lambda,\mu},\quad |z|=1,\quad \lambda,\mu=0,1. \tag{3.2}$$

We also remark that if P^λ, $G^\lambda \in \mathcal{W}$ satisfy (3.1) or (3.2), then there exists some $R\in\mathcal{W}$, with $R(z)\neq 0$ on $|z|=1$, such that

$$\begin{aligned}&P^0(z)G^0(z)+P^0(-z)G^0(-z)=1,\\ &P^1(z)=-zG^0(-z)R(z^2),\quad G^1(z)=\frac{-z^{-1}P^0(-z)}{R(z^2)},\quad |z|=1.\end{aligned} \tag{3.3}$$

That is, we have the following result (see [1, p. 148]).

LEMMA 1. *Let P^λ and G^λ, $\lambda=0,1$, be Laurent series in the class $\mathcal{W}$. Then the three statements* (3.1), (3.2), *and* (3.3) *are equivalent.*

Proof. That (3.3) implies (3.2) is trivial. On the other hand, to see that (3.2) implies (3.1), we will use the notation introduced in (2.11). It is clear that (3.2) is equivalent to the matrix identity

$$M^T(z)\widetilde{M}(z)=I,\qquad |z|=1, \tag{3.4}$$

where I denotes the identity matrix and A^T stands for the transpose of A. Hence, $M(z)$ is a nonsingular matrix for $|z|=1$ with inverse $\widetilde{M^T}(z)$, and we may also conclude that

$$M(z)\widetilde{M^T}(z)=I,\qquad |z|=1. \tag{3.5}$$

Multiplying out the matrices in (3.5), we see that two of the four identities constitute (3.1).

Finally, suppose that (3.1) holds. Replacing z by $-z$, we have

$$\begin{aligned}&P^0(z)G^0(-z)+P^1(z)G^1(-z)=0,\\ &P^0(-z)G^0(-z)+P^1(-z)G^1(-z)=1,\qquad |z|=1.\end{aligned} \tag{3.6}$$

Then (3.5) is a consequence of (3.1) and (3.6). In particular, the matrix $M(z)$ is nonsingular for $|z|=1$. Viewing $G^0(z)$ and $G^1(z)$ as unknowns in (3.1), we apply Cramer's rule to yield

$$\begin{aligned}G^0(z)&=\frac{P^1(-z)}{\Delta(z)};\\ G^1(z)&=-\frac{P^0(-z)}{\Delta(z)},\qquad |z|=1,\end{aligned} \tag{3.7}$$

where

$$\Delta(z):=\det M(z). \tag{3.8}$$

From the definition of $M(z)$ and $\Delta(z)$, it is clear that $\Delta(z)$ is an odd function in z; and hence, we may write

$$\Delta(z)=zR(z^2),\qquad |z|=1 \tag{3.9}$$

for some $R \in \mathcal{W}$. Since $M(z)$ is nonsingular on $|z| = 1$, we have $R(z) \neq 0$, $|z| = 1$. From (3.7) and (3.9) as well as the equivalence of (3.4) and (3.5), we establish (3.3). □

In what follows, we will always assume that P^0 and G^0 are dual symbols as in (2.9), that P^1 and G^1 are defined by (2.10), and that the series expressions (2.6) and (2.12) are used. With these (two-scale) coefficient sequences $\{p_n^\lambda\}$ and $\{\overline{g_{-n}^\lambda}\}$, $\lambda = 0, 1$, we introduce the linear operators $\mathcal{P}_\lambda$ and $\mathcal{G}_\lambda$, $\lambda = 0, 1$, on

$$\ell^2 := \left\{ \mathbf{v} = \{v_k\} \colon \sum_{k\in\mathbf{Z}} |v_k|^2 < \infty \right\}, \tag{3.10}$$

defined by

$$(\mathcal{P}_\lambda \mathbf{v})_\ell := \sum_{k\in\mathbf{Z}} p_{k-2\ell}^\lambda v_k, \quad \ell \in \mathbb{Z}, \quad \lambda = 0, 1, \tag{3.11}$$

$$(\mathcal{G}_\lambda \mathbf{v})_\ell := \sum_{k\in\mathbf{Z}} \overline{g_{-k+2\ell}^\lambda} v_k, \quad \ell \in \mathbb{Z}, \quad \lambda = 0, 1. \tag{3.12}$$

Then the adjoints $\mathcal{P}_\lambda^*$ and $\mathcal{G}_\lambda^*$ of $\mathcal{P}_\lambda$ and $\mathcal{G}_\lambda$, respectively, are given by

$$(\mathcal{P}_\lambda^* \mathbf{v})_\ell = \sum_{k\in\mathbf{Z}} \overline{p_{\ell-2k}^\lambda} v_k, \quad \ell \in \mathbb{Z}, \quad \lambda = 0, 1, \tag{3.13}$$

$$(\mathcal{G}_\lambda^* \mathbf{v})_\ell = \sum_{k\in\mathbf{Z}} g_{-\ell+2k}^\lambda v_k, \quad \ell \in \mathbb{Z}, \quad \lambda = 0, 1. \tag{3.14}$$

Indeed, by adopting the notation

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\ell^2} = \sum_{k\in\mathbf{Z}} u_k \bar{v}_k, \tag{3.15}$$

where $\mathbf{u} = \{u_k\}$ and $\mathbf{v} = \{v_k\}$, for the inner product of the ℓ^2 space, we have

$$\begin{aligned} \langle \mathbf{u}, \mathcal{P}_\lambda \mathbf{v} \rangle &= \sum_{\ell\in\mathbf{Z}} u_\ell \overline{(\mathcal{P}_\lambda \mathbf{v})_\ell} \\ &= \sum_{k\in\mathbf{Z}} \left(\sum_{\ell\in\mathbf{Z}} \overline{p_{k-2\ell}^\lambda} u_\ell \right) \overline{v}_k = \langle \mathcal{P}_\lambda^* \mathbf{u}, \mathbf{v} \rangle, \end{aligned} \tag{3.16}$$

and similarly,

$$\langle \mathbf{u}, \mathcal{G}_\lambda \mathbf{v} \rangle = \langle \mathcal{G}_\lambda^* \mathbf{u}, \mathbf{v} \rangle. \tag{3.17}$$

We have the following result.

LEMMA 2. *Let $\mathcal{P}_\lambda$ and $\mathcal{G}_\lambda$ be as defined in* (3.11) *and* (3.12) *with adjoints $\mathcal{P}_\lambda^*$ and $\mathcal{G}_\lambda^*$, respectively. Then*

$$\mathcal{P}_\lambda \mathcal{G}_\mu^* = 2\delta_{\lambda,\mu} I, \qquad \lambda, \mu = 0, 1, \tag{3.18}$$

and

$$\mathcal{G}_0^* \mathcal{P}_0 + \mathcal{G}_1^* \mathcal{P}_1 = 2I, \tag{3.19}$$

where I is the identity operator on ℓ^2.

Remarks. (1) By taking the adjoints of both (3.18) and (3.19), we also have

$$\mathcal{G}_\mu \mathcal{P}_\lambda^* = 2\delta_{\lambda,\mu} I, \qquad \lambda, \mu = 0, 1, \tag{3.20}$$

and

$$\mathcal{P}_0^* \mathcal{G}_0 + \mathcal{P}_1^* \mathcal{G}_1 = 2I. \tag{3.21}$$

That is, the identities (3.18) and (3.19) are equivalent to (3.20) and (3.21), respectively.

(2) We can view $\mathcal{P}_\lambda$ and $\mathcal{G}_\lambda$, $\lambda = 0, 1$, as bi-infinite (also possibly finite) matrices, namely,

$$\mathcal{P}_\lambda = [p^\lambda_{k-2\ell}]_{\ell,k\in\mathbb{Z}}, \qquad \mathcal{G}_\lambda = [\overline{g^\lambda_{k-2\ell}}]_{\ell,k\in\mathbb{Z}}.$$

With $\mathcal{P}_\lambda^* := \overline{\mathcal{P}}_\lambda^T$ and $\mathcal{G}_\lambda^* := \overline{\mathcal{G}}_\lambda^T$ being the conjugate transposes of $\mathcal{P}_\lambda$ and $\mathcal{G}_\lambda$, respectively, and I being the identity matrix, then (3.18)–(3.21) can be viewed as matrix identities.

Proof. We will derive (3.18) and (3.19) via symbol calculus. For any sequence $\mathbf{u} = \{u_k\}_{k\in\mathbb{Z}} \in \ell^1$, set

$$\mathbf{u}(z) := \sum_{k\in\mathbb{Z}} u_k z^k.$$

According to (3.11) and (3.14), it is easy to verify that

$$(\mathcal{P}_\lambda \mathbf{v})(z^{-2}) = \sum_{\ell\in\mathbb{Z}} (\mathcal{P}_\lambda \mathbf{v})_\ell z^{-2\ell} = P^\lambda(z)\mathbf{v}(z^{-1}) + P^\lambda(-z)\mathbf{v}(-z^{-1})$$

and

$$(\mathcal{G}_\lambda^* \mathbf{v})(z^{-1}) = \sum_{\ell\in\mathbb{Z}} (\mathcal{G}_\lambda^* \mathbf{v})_\ell z^{-\ell} = 2G^\lambda(z)\mathbf{v}(z^{-2}).$$

Thus, we have, for any $\mathbf{u} \in \ell^1$ and $|z| = 1$,

$$\begin{aligned}\sum_{\ell\in\mathbb{Z}} (\mathcal{P}_\lambda \mathcal{G}_\mu^* \mathbf{u})_\ell z^{-2\ell} =& P^\lambda(z)(\mathcal{G}_\mu^* \mathbf{u})(z^{-1}) + P^\lambda(-z)(\mathcal{G}_\mu^* \mathbf{u})(-z^{-1}) \\ =& 2[P^\lambda(z)G^\mu(z) + P^\lambda(-z)G^\mu(-z)]\mathbf{u}(z^{-2}) = 2\delta_{\lambda,\mu}\mathbf{u}(z^{-2}),\end{aligned}$$

where the identity (3.2) has been used to arrive at the last equality. That is, we have $(\mathcal{P}_\lambda \mathcal{G}_\mu^* \mathbf{u})_k = 2\delta_{\lambda,\mu} u_k$ for all $k \in \mathbb{Z}$ and all $\mathbf{u} = \{u_k\}_{k\in\mathbb{Z}} \in \ell^1$. Hence, since ℓ^1 is dense in ℓ^2, we obtain (3.18).

On the other hand, we have, for any $\mathbf{v} \in \ell^1$,

$$\begin{aligned}&\sum_{k\in\mathbb{Z}} (\mathcal{G}_0^* \mathcal{P}_0 \mathbf{v} + \mathcal{G}_1^* \mathcal{P}_1 \mathbf{v})_k z^{-k} = (\mathcal{G}_0^* \mathcal{P}_0 \mathbf{v})(z^{-1}) + (\mathcal{G}_1^* \mathcal{P}_1 \mathbf{v})(z^{-1}) \\ &=2G^0(z)(\mathcal{P}_0 \mathbf{v})(z^{-2}) + 2G^1(z)(\mathcal{P}_1 \mathbf{v})(z^{-2}) \\ &=2G^0(z)[P^0(z)\mathbf{v}(z^{-1}) + P^0(-z)\mathbf{v}(-z^{-1})] + 2G^1(z)[P^1(z)\mathbf{v}(z^{-1}) + P^1(-z)\mathbf{v}(-z^{-1})] \\ &=2[P^0(z)G^0(z) + P^1(z)G^1(z)]\mathbf{v}(z^{-1}) + 2[P^0(-z)G^0(z) + P^1(-z)G^1(z)]\mathbf{v}(-z^{-1}) \\ &=2\mathbf{v}(z^{-1}),\end{aligned}$$

where the identity (3.1) has been used to arrive at the last equality. Thus,

$$(\mathcal{G}_0^* \mathcal{P}_0 \mathbf{v} + \mathcal{G}_1^* \mathcal{P}_1 \mathbf{v})_k = 2v_k \quad \text{for all} \quad k \in \mathbb{Z}.$$

Since this holds for all $\mathbf{v} \in \ell^1$, we obtain (3.19). □

As a consequence of Lemma 2, we see that

$$\left(\tfrac{1}{2}\mathcal{G}_\lambda^*\mathcal{P}_\lambda\right)\left(\tfrac{1}{2}\mathcal{G}_\lambda^*\mathcal{P}_\lambda\right) = \tfrac{1}{2}\mathcal{G}_\lambda^*\left(\tfrac{1}{2}\mathcal{P}_\lambda\mathcal{G}_\lambda^*\right)\mathcal{P}_\lambda = \tfrac{1}{2}\mathcal{G}_\lambda^*\mathcal{P}_\lambda,$$

and

$$\left(\tfrac{1}{2}\mathcal{G}_\lambda^*\mathcal{P}_\lambda\right)\left(\tfrac{1}{2}\mathcal{G}_\mu^*\mathcal{P}_\mu\right) = \tfrac{1}{2}\mathcal{G}_\lambda^*\left(\tfrac{1}{2}\mathcal{P}_\lambda\mathcal{G}_\mu^*\right)\mathcal{P}_\mu = 0, \qquad \lambda \neq \mu.$$

This shows that

(i) $\frac{1}{2}\mathcal{G}_\lambda^*\mathcal{P}_\lambda$, is a projection on ℓ^2, $\lambda = 0, 1$;

(ii) The ranges of $\frac{1}{2}\mathcal{P}_0^*\mathcal{G}_0$ and $\frac{1}{2}\mathcal{G}_1^*\mathcal{P}_1$ are orthogonal (since $\langle \mathcal{P}_0^*\mathcal{G}_0\mathbf{u}, \mathcal{G}_1^*\mathcal{P}_1\mathbf{v}\rangle_{\ell^2} = \langle \mathbf{u}, \mathcal{G}_0^*\mathcal{P}_0\mathcal{G}_1^*\mathcal{P}_1\mathbf{v}\rangle_{\ell^2} = 0$); while

(iii) The ranges of $\frac{1}{2}\mathcal{G}_0^*\mathcal{P}_0$ and $\frac{1}{2}\mathcal{G}_1^*\mathcal{P}_1$ form a direct-sum decomposition of ℓ^2.

We now return to our discussion of wavelet packets as defined in §2 (cf. (2.14)). From (2.14) and using the notation in (3.11), we have

$$\psi_{2n+\lambda}(x-\ell) = (\mathcal{P}_\lambda\{\psi_n(2x-\cdot)\})_\ell, \qquad \ell \in \mathbb{Z}, \quad \lambda = 0, 1. \tag{3.22}$$

Similarly, we also have, from (2.14) and (3.12),

$$\widetilde{\psi}_{2n+\lambda}(x-\ell) = (\mathcal{G}_\lambda\{\widetilde{\psi}_n(2x-\cdot)\})_\ell, \qquad \ell \in \mathbb{Z}, \quad \lambda = 0, 1. \tag{3.23}$$

As an application of (3.19) in Lemma 2, by putting (3.22) into (3.14), we obtain, for all $k \in \mathbb{Z}$,

$$\begin{aligned} &\psi_n(2x-k) \\ &\quad= \frac{1}{2}(\mathcal{G}_0^*(\mathcal{P}_0\{\psi_n(2x-\cdot)\}))_k + \frac{1}{2}(\mathcal{G}_1^*(\mathcal{P}_1\{\psi_n(2x-\cdot)\}))_k \\ &\quad= \frac{1}{2}(\mathcal{G}_0^*\{\psi_{2n}(x-\cdot)\})_k + \frac{1}{2}(\mathcal{G}_1^*\{\psi_{2n+1}(x-\cdot)\})_k \\ &\quad= \frac{1}{2}\sum_{\ell\in\mathbf{Z}} g_{2\ell-k}^0\psi_{2n}(x-\ell) + \frac{1}{2}\sum_{\ell\in\mathbf{Z}} g_{2\ell-k}^1\psi_{2n+1}(x-\ell). \end{aligned} \tag{3.24}$$

In the same manner, from (3.23), (3.21), and (3.13), we also have, for $k \in \mathbb{Z}$,

$$\begin{aligned} &\widetilde{\psi}_n(2x-k) \\ &\quad= \frac{1}{2}(\mathcal{P}_0^*(\mathcal{G}_0\{\widetilde{\psi}_n(2x-\cdot)\}))_k + \frac{1}{2}(\mathcal{P}_1^*(\mathcal{G}_1\{\widetilde{\psi}_n(2x-\cdot)\}))_k \\ &\quad= \frac{1}{2}(\mathcal{P}_0^*\{\widetilde{\psi}_{2n}(x-\cdot)\})_k + \frac{1}{2}(\mathcal{P}_1^*\{\widetilde{\psi}_{2n+1}(x-\cdot)\})_k \\ &\quad= \frac{1}{2}\sum_{\ell\in\mathbf{Z}} \overline{p_{k-2\ell}^0}\,\widetilde{\psi}_{2n}(x-\ell) + \frac{1}{2}\sum_{\ell\in\mathbf{Z}} \overline{p_{k-2\ell}^1}\,\widetilde{\psi}_{2n+1}(x-\ell). \end{aligned} \tag{3.25}$$

Observe that for $n = 0$, (3.24) and (3.25) give the so-called decomposition relations in wavelet analysis [1, Chap. 5]. In general, we call (3.24) and (3.25) "*decomposition formulas*" for the wavelet packets $\{\psi_n\}$ and $\{\widetilde{\psi}_n\}$, $n \in \mathbb{Z}_+$, respectively, and (3.19), (3.21) the "*operator forms*" of these decomposition formulas.

In the following, we give explicit expressions of the Fourier transforms of ψ_n and $\widetilde{\psi}_n$.

LEMMA 3. *Let $n \in \mathbb{Z}_+$ and consider its dyadic expansion*

$$n = \sum_{j=1}^{\infty} \varepsilon_j 2^{j-1}, \qquad \varepsilon_j \in \{0, 1\}. \tag{3.26}$$

Then

$$\widehat{\psi}_n(\omega) = \prod_{J=1}^{\infty} P^{\varepsilon_j}(e^{-i2^{-j}\omega}), \tag{3.27}$$

$$\widehat{\widetilde{\psi}}_n(\omega) = \prod_{j=1}^{\infty} \overline{G^{\varepsilon_j}(e^{-i2^{-j}\omega})}. \tag{3.28}$$

Remark. Since n is an integer, we note that all but finitely many ε_j in (3.26) are zeros.

Proof. We will prove (3.27) by induction. The proof of (3.28) is identical. From (2.4), noting that $\psi_0 = \phi$, we have (3.27) for $n = 0$. Suppose now that (3.27) is valid for all n, $0 \le n < 2^N$, where N is a nonnegative integer. Then for $2^N \le n < 2^{N+1}$ with the dyadic expansion (3.26), we have

$$\left\lfloor \frac{n}{2} \right\rfloor = \sum_{j=1}^{\infty} \varepsilon_{j+1} 2^{j-1} \le \frac{n}{2} < 2^N$$

and

$$n = 2\left\lfloor \frac{n}{2} \right\rfloor + \varepsilon_1.$$

Thus, from the Fourier transform formulation of (2.14), namely,

$$\widehat{\psi}_{2n+\lambda}(\omega) = P^{\lambda}(e^{-i\omega/2})\widehat{\psi}_n\left(\frac{\omega}{2}\right),$$

and applying the induction hypothesis, we obtain

$$\begin{aligned} \widehat{\psi}_n(\omega) &= P^{\varepsilon_1}(e^{-i\omega/2})\widehat{\psi}_{\lfloor \frac{n}{2} \rfloor}\left(\frac{\omega}{2}\right) \\ &= P^{\varepsilon_1}(e^{-i\omega/2}) \prod_{j=1}^{\infty} P^{\varepsilon_{j+1}}(e^{-i2^{-j-1}\omega}) = \prod_{j=1}^{\infty} P^{\varepsilon_j}(e^{-i2^{-j}\omega}). \qquad \square \end{aligned}$$

We next discuss the duality properties between the wavelet packets $\{\psi_n\}_{n=1}^{\infty}$ and $\{\widetilde{\psi}_n\}_{n=1}^{\infty}$.

LEMMA 4. *For all $k, \ell \in \mathbb{Z}$, and $n \in \mathbb{Z}_+$,*

$$\langle \psi_n(\cdot - k), \widetilde{\psi}_n(\cdot - \ell) \rangle = \delta_{k,\ell}. \tag{3.29}$$

Proof. We will establish (3.29) by induction on n. The case $n = 0$ is our basic assumption (2.8) on the dual scaling functions $\psi_0 = \phi$ and $\widetilde{\psi}_0 = \widetilde{\phi}$. Suppose that (3.29) holds for $0 \le n < 2^N$, where N is a nonnegative integer. Then for $2^N \le n < 2^{N+1}$, since we can write $n = 2\lfloor \frac{n}{2} \rfloor + \lambda$ for some $\lambda \in \{0, 1\}$ according to the proof of Lemma 3, we have, from the Fourier transform formulations of both equations in (2.14),

$$\begin{aligned}
&\langle \psi_n(\cdot - k), \widetilde{\psi}_n(\cdot - \ell) \rangle \\
&= \frac{1}{2\pi} \int_{\mathbf{R}} \widehat{\psi}_n(\omega) \overline{\widehat{\widetilde{\psi}}_n(\omega)}\, e^{i(\ell - k)\omega}\, d\omega \\
&= \frac{1}{2\pi} \int_{\mathbf{R}} P^{\lambda}(e^{-i\omega/2}) G^{\lambda}(e^{-i\omega/2}) \widehat{\psi}_{\lfloor \frac{n}{2} \rfloor}\left(\frac{\omega}{2}\right) \overline{\widehat{\widetilde{\psi}}_{\lfloor \frac{n}{2} \rfloor}\left(\frac{\omega}{2}\right)}\, e^{i(\ell - k)\omega}\, d\omega \\
&= \frac{1}{2\pi} \int_0^{4\pi} e^{i(\ell - k)\omega} P^{\lambda}(e^{-i\omega/2}) G^{\lambda}(e^{-i\omega/2}) \sum_{j \in \mathbf{Z}} \widehat{\psi}_{\lfloor \frac{n}{2} \rfloor}\left(\frac{\omega}{2} + 2\pi j\right) \overline{\widehat{\widetilde{\psi}}_{\lfloor \frac{n}{2} \rfloor}\left(\frac{\omega}{2} + 2\pi j\right)}\, d\omega.
\end{aligned}$$

Since $\lfloor \frac{n}{2} \rfloor \le \frac{n}{2} < 2^N$, it follows from the induction hypothesis that $\langle \psi_{\lfloor \frac{n}{2} \rfloor}(\cdot - k), \widetilde{\psi}_{\lfloor \frac{n}{2} \rfloor}(\cdot - \ell) \rangle = \delta_{k,\ell}$ for all $k, \ell \in \mathbb{Z}$, and this is equivalent to

$$\sum_{j \in \mathbf{Z}} \widehat{\psi}_{\lfloor \frac{n}{2} \rfloor}\left(\frac{\omega}{2} + 2\pi j\right) \overline{\widehat{\widetilde{\psi}}_{\lfloor \frac{n}{2} \rfloor}\left(\frac{\omega}{2} + 2\pi j\right)} = 1 \quad \text{a.e.,} \tag{3.30}$$

via the Poisson summation formula (see [1, pp. 45–46 and pp. 151–152]). Thus, we have, from (3.2) for $\mu = \lambda$ that

$$\begin{aligned}
&\langle \psi_n(\cdot - k), \widetilde{\psi}_n(\cdot - \ell) \rangle \\
&= \frac{1}{2\pi} \int_0^{4\pi} e^{i(\ell - k)\omega} P^{\lambda}(e^{-i\omega/2}) G^{\lambda}(e^{-i\omega/2})\, d\omega \\
&= \frac{1}{2\pi} \int_0^{2\pi} e^{i(\ell - k)\omega} [P^{\lambda}(e^{-i\omega/2}) G^{\lambda}(e^{-i\omega/2}) + P^{\lambda}(-e^{-i\omega/2}) G^{\lambda}(-e^{-i\omega/2})]\, d\omega \\
&= \frac{1}{2\pi} \int_0^{2\pi} e^{i(\ell - k)\omega}\, d\omega = \delta_{k,\ell}, \qquad k, \ell \in \mathbb{Z}.
\end{aligned}$$

This shows that (3.29) also holds for $2^N \le n < 2^{N+1}$. □

LEMMA 5. *For all* $k, \ell \in \mathbb{Z}$, $n \in \mathbb{Z}_+$, *and* $\lambda, \mu \in \{0, 1\}$, *with* $\lambda \ne \mu$,

$$\langle \psi_{2n+\lambda}(\cdot - k), \widetilde{\psi}_{2n+\mu}(\cdot - \ell) \rangle = 0. \tag{3.31}$$

Proof. By applying the Fourier transform formulations of (2.14) and the formulas (3.30) and (3.2), we have, as in the proof of Lemma 4, that

$$\begin{aligned}
&\langle \psi_{2n+\lambda}(\cdot - k), \widetilde{\psi}_{2n+\mu}(\cdot - \ell) \rangle \\
&= \frac{1}{2\pi} \int_{\mathbf{R}} \widehat{\psi}_{2n+\lambda}(\omega) \overline{\widehat{\widetilde{\psi}}_{2n+\mu}(\omega)}\, e^{i(\ell - k)\omega}\, d\omega \\
&= \frac{1}{2\pi} \int_{\mathbf{R}} P^{\lambda}(e^{-i\omega/2}) G^{\mu}(e^{-i\omega/2}) \widehat{\psi}_n\left(\frac{\omega}{2}\right) \overline{\widehat{\widetilde{\psi}}_n\left(\frac{\omega}{2}\right)} e^{i(\ell - k)\omega}\, d\omega \\
&= \frac{1}{2\pi} \int_0^{4\pi} e^{i(\ell - k)\omega} P^{\lambda}(e^{-i\omega/2}) G^{\mu}(e^{-i\omega/2}) \sum_{j \in \mathbf{Z}} \widehat{\psi}_n\left(\frac{\omega}{2} + 2\pi j\right) \overline{\widehat{\widetilde{\psi}}_n\left(\frac{\omega}{2} + 2\pi j\right)}\, d\omega \\
&= \frac{1}{2\pi} \int_0^{4\pi} e^{i(\ell - k)\omega} P^{\lambda}(e^{-i\omega/2}) G^{\mu}(e^{-i\omega/2})\, d\omega \\
&= \frac{1}{2\pi} \int_0^{2\pi} e^{i(\ell - k)\omega} [P^{\lambda}(e^{-i\omega/2}) G^{\mu}(e^{-i\omega/2}) + P^{\lambda}(-e^{i\omega/2}) G^{\mu}(-e^{-i\omega/2})]\, d\omega \\
&= \frac{1}{2\pi} \int_0^{2\pi} e^{i(\ell - k)\omega} \delta_{\lambda,\mu} d\omega = \delta_{\lambda,\mu} \delta_{k,\ell} = 0, \qquad \lambda \ne \mu.
\end{aligned}$$

This completes the proof of Lemma 5. □

Recall the spaces U_n of wavelet packets as defined in (2.24). Any f in U_n can be written as

$$f(x) = \sum_{k\in\mathbf{Z}} \bar{c}_k \psi_n(x-k) = \langle \{\psi_n(x-\cdot)\}, \mathbf{c}\rangle_{\ell^2}, \tag{3.32}$$

for some $\mathbf{c} = \{c_k\}_{k\in\mathbf{Z}} \in \ell^2$. From the decomposition formula (3.24), we have

$$\begin{aligned} f(2x) &= \langle \{\psi_n(2x-\cdot)\}, \mathbf{c}\rangle_{\ell^2} \\ &= \frac{1}{2}\langle \mathcal{G}_0^*\{\psi_{2n}(x-\cdot)\}, \mathbf{c}\rangle_{\ell^2} + \frac{1}{2}\langle \mathcal{G}_1^*\{\psi_{2n+1}(x-\cdot)\}, \mathbf{c}\rangle_{\ell^2} \\ &= \frac{1}{2}\langle \{\psi_{2n}(x-\cdot)\}, \mathcal{G}_0\mathbf{c}\rangle_{\ell^2} + \frac{1}{2}\langle \{\psi_{2n+1}(x-\cdot)\}, \mathcal{G}_1\mathbf{c}\rangle_{\ell^2} \\ &= \frac{1}{2}\sum_{\ell\in\mathbf{Z}} \overline{(\mathcal{G}_0\mathbf{c})_\ell}\, \psi_{2n}(x-\ell) + \frac{1}{2}\sum_{\ell\in\mathbf{Z}} \overline{(\mathcal{G}_1\mathbf{c})_\ell}\, \psi_{2n+1}(x-\ell). \end{aligned} \tag{3.33}$$

By introducing the scaling operator

$$\delta f := f(2\cdot), \tag{3.34}$$

the formula in (3.33) implies that for $f \in U_n$, we have

$$\delta f = g + h, \tag{3.35}$$

where $g \in U_{2n}$, $h \in U_{2n+1}$. On the other hand, if $g = \sum_{\ell\in\mathbf{Z}} a_\ell \psi_{2n}(\cdot-\ell) \in U_{2n}$ and $h = \sum_{\ell\in\mathbf{Z}} b_\ell \psi_{2n+1}(\cdot-\ell) \in U_{2n+1}$ such that

$$g + h = 0, \tag{3.36}$$

then by Lemmas 4 and 5 we have

$$\begin{aligned} 0 &= \langle g+h, \tilde{\psi}_{2n}(\cdot-k)\rangle = a_k, \\ 0 &= \langle g+h, \tilde{\psi}_{2n+1}(\cdot-k)\rangle = b_k, \qquad k \in \mathbb{Z}. \end{aligned}$$

That is, we have $g = 0$ and $h = 0$. Noting from the definition in (2.14) that

$$U_{2n} \subseteq \delta U_n, \qquad U_{2n+1} \subseteq \delta U_n, \tag{3.37}$$

the above argument shows that

$$\delta U_n = U_{2n} \dot{+} U_{2n+1}, \qquad n \in \mathbb{Z}_+, \tag{3.38}$$

where "$\dot{+}$" indicates a "direct sum" of two linear spaces. Analogous results also hold for $\widetilde{U}_n$, namely,

$$\delta \widetilde{U}_n = \widetilde{U}_{2n} \dot{+} \widetilde{U}_{2n+1}, \qquad n \in \mathbb{Z}_+. \tag{3.39}$$

More generally, we have

$$\begin{aligned} \delta^\ell U_n &= U_{2^\ell n} \dot{+} U_{2^\ell n+1} \dot{+} \cdots \dot{+} U_{2^\ell n + 2^\ell - 1}, \\ \delta^\ell \widetilde{U}_n &= \widetilde{U}_{2^\ell n} \dot{+} \widetilde{U}_{2^\ell n+1} \dot{+} \cdots \dot{+} \widetilde{U}_{2^\ell n + 2^\ell - 1}, \end{aligned} \tag{3.40}$$

where $\ell, n \in \mathbb{Z}_+$, $\delta^0 := I$, and $\delta^\ell = \delta \cdot \delta^{\ell-1}$. Of course, the proof of (3.40) is simply repeated applications of (3.38) and (3.39). In addition, an equivalent statement of Lemma 5 is that

$$(3.41) \qquad U_{2n} \perp \tilde{U}_{2n+1} \quad \text{and} \quad U_{2n+1} \perp \tilde{U}_{2n}, \qquad n \in \mathbb{Z}_+.$$

We end this section by showing that the infinite direct-sums in (2.23) are indeed "direct-sum decompositions." More precisely, for any $f \in L^2$, we will show that there exist unique $g_j \in W_j$, $j \in \mathbb{Z}$, such that

$$(3.42) \qquad \lim_{m,n\to\infty} \left\| \sum_{j=-m}^{n} g_j - f \right\| = 0.$$

Consider the linear operators P_j and Δ_j, defined by

$$(3.43) \qquad P_j f := \sum_{k\in\mathbb{Z}} \langle f, \tilde{\phi}_{j,k} \rangle \phi_{j,k} \quad \text{and} \quad \Delta_j f := \sum_{k\in\mathbb{Z}} \langle f, \tilde{\psi}_{j,k} \rangle \psi_{j,k}.$$

It is clear that $P_{j+1} = P_j + \Delta_j$ for any $j \in \mathbb{Z}$, so that

$$(3.44) \qquad P_{n+1} = P_{-m} + \sum_{j=-m}^{n} \Delta_j, \qquad m,\ n \in \mathbb{Z}_+.$$

Let $f \in L^2$ and $\varepsilon > 0$ be arbitrarily given. Then by the definition of the MRA $\{V_j\}$, there exist some $J \in \mathbb{Z}_+$ and an $f_J \in V_J$ such that $\|f - f_J\| < \varepsilon$. Since $P_j(f_J) = f_J$ for all $j \geq J$, we have

$$\begin{aligned} \|P_j f - f\| &\leq \|P_j(f - f_J)\| + \|f_J - f\| \\ &\leq (\|P_j\| + 1)\varepsilon, \qquad j \geq J; \end{aligned}$$

and the uniform boundedness of $\|P_j\|$ (in fact, $\|P_j\| = \|P_0\|$ for all $j \in \mathbb{Z}$) now yields

$$(3.45) \qquad \lim_{j\to\infty} \|P_j f - f\| = 0.$$

Hence, it follows from (3.44), (3.45), and $\lim_{j\to-\infty} \|P_j f\| = 0$ (which is a consequence of $\cap V_j = \{0\}$), that

$$\left\| \sum_{j=-m}^{n} \Delta_j f - f \right\| \leq \|P_{n+1} f - f\| + \|P_{-m} f\| \to 0 \quad \text{as} \ m,\ n \to +\infty.$$

Since $\Delta_j f \in W_j$, we have established the existence of $g_j \in W_j$ in (3.42). To show the uniqueness of g_j, $j \in \mathbb{Z}$, we first observe from the duality property in (1.4) that

$$(3.46) \qquad \Delta_j \left(\sum_{\ell=-J}^{J} g_\ell \right) = \Delta_j g_j = g_j, \qquad |j| \leq J \in \mathbb{Z}_+.$$

Hence, for any fixed $j \in \mathbb{Z}$, we have, from (3.42) and (3.46),

$$0 = \lim_{J\to\infty} \left\| \Delta_j \left(\sum_{\ell=-J}^{J} g_\ell - f \right) \right\| = \|g_j - \Delta_j f\|.$$

That is, $g_j = \Delta_j f$. □

4. Proof of the main results. We are now ready to prove the theorems stated in §2.

Proof of Theorem 1. In view of Lemma 4, we only have to consider the case $m \neq n$. We first note that

$$U_n \subseteq \delta^t U_{\lfloor \frac{n}{2^t} \rfloor}, \qquad n, t \in \mathbb{Z}_+, \tag{4.1}$$

where as before, $\lfloor x \rfloor$ stands for the largest integer not exceeding x. Indeed, by observing that $n \in \left[2^t \lfloor \frac{n}{2^t} \rfloor, 2^t \left(\lfloor \frac{n}{2^t} \rfloor + 1\right)\right)$, it follows from (3.40) that

$$U_n \subseteq U_{2^t\lfloor\frac{n}{2^t}\rfloor} \dot{+} U_{2^t\lfloor\frac{n}{2^t}\rfloor+1} \dot{+} \cdots \dot{+} U_{2^t\lfloor\frac{n}{2^t}\rfloor+2^t-1} = \delta^t U_{\lfloor\frac{n}{2^t}\rfloor}.$$

Similarly, we also have

$$\widetilde{U}_n \subseteq \delta^t \widetilde{U}_{\lfloor \frac{n}{2^t} \rfloor}, \qquad n, t \in \mathbb{Z}_+. \tag{4.2}$$

Now, without loss of generality, let us assume that $m > n$ in (2.17). Using the dyadic expansions for both m and n, namely,

$$m = \sum_{j\in\mathbb{Z}_+} 2^j \varepsilon_j, \quad n = \sum_{j\in\mathbb{Z}_+} 2^j \varepsilon'_j, \quad \varepsilon_j, \varepsilon'_j \in \{0, 1\},$$

we can find some $t \in \mathbb{Z}_+$ such that $\varepsilon'_j = \varepsilon_j$ for all $j > t$, while $\varepsilon_t > \varepsilon'_t$; i.e., $\varepsilon_t = 1$, $\varepsilon'_t = 0$. Thus, by letting $n' := \sum_{j>t} 2^{j-t-1}\varepsilon_j$, we have $n' \in \mathbb{Z}_+$ and $\lfloor \frac{m}{2^t} \rfloor = 2n' + 1$, $\lfloor \frac{n}{2^t} \rfloor = 2n'$. Hence, it follows from (4.1) and (4.2) that $\psi_m(\cdot - k) \in U_m \subseteq \delta^t U_{2n'+1}$ and $\widetilde{\psi}_n(\cdot - \ell) \in \widetilde{U}_n \subseteq \delta^t \widetilde{U}_{2n'}$ for all $k, \ell \in \mathbb{Z}$, respectively. On the other hand, from (3.41), we have shown that $\delta^t U_{2n'+1} \perp \delta^t \widetilde{U}_{2n'}$ and therefore $\langle \psi_m(\cdot - k), \psi_n(\cdot - \ell)\rangle = 0$. This completes the proof of Theorem 1. □

Remark. In the above discussion, we have proved that for any integers $m > n \geq 0$, there exists a $t \in \mathbb{Z}_+$, such that $\lfloor \frac{m}{2^t} \rfloor - \lfloor \frac{n}{2^t} \rfloor = 1$, while $\lfloor \frac{m}{2^t} \rfloor$ is an odd integer and $\lfloor \frac{n}{2^t} \rfloor$ is an even integer.

Proof of Theorem 2. Noting that $\psi_0(\cdot - k) = \psi_{0;0,k}$ and $\psi_1(\cdot - k) = \psi_{1;0,k}$, in (2.18), (2.19), we see that $V_0 = U_0$, $W_0 = U_1$, $V_j = \delta^j V_0$ and $W_j = \delta^j W_0$. Thus, as an application of formula (3.40) for $\ell = j$ and $n = 0, 1$, we immediately obtain (2.25) for any $j \in \mathbb{Z}_+$. The proof of (2.26) is the same. Since each of $\{V_j\}$ and $\{\widetilde{V}_j\}$ forms a multiresolution analysis of L^2, it is clear that

$$L^2 = V_0 \dot{+} \overset{\bullet}{\sum_{j\in\mathbb{Z}_+}} W_j = \widetilde{V}_0 \dot{+} \overset{\bullet}{\sum_{j\in\mathbb{Z}_+}} \widetilde{W}_j.$$

Combining this fact with (2.25) and (2.26) gives (2.27). Finally, (2.28) is a consequence of Theorem 1. □

Proof of Theorem 3. By Theorems 1 and 2, we see that both of the families

$$\{\psi_n(\cdot - k)\colon\ k \in \mathbb{Z},\ 2^\ell \leq n < 2^{\ell+1}\} \quad \text{and} \quad \{\widetilde{\psi}_n(\cdot - k)\colon\ k \in \mathbb{Z},\ 2^\ell \leq n < 2^{\ell+1}\}$$

are bases of W_ℓ and $\widetilde{W}_\ell$, respectively, and are biorthogonal to each other. Thus, by observing the definitions (2.24) and (3.34), we see that

$$\{\psi_{n;j,k}\colon\ k \in \mathbb{Z},\ 2^\ell \leq n < 2^{\ell+1}\} \quad \text{and} \quad \{\widetilde{\psi}_{n;j',k}\colon\ k \in \mathbb{Z},\ 2^\ell \leq n < 2^{\ell+1}\}$$

are bases of $W_{\ell+j} = \delta^j W_\ell$ and $W_{\ell+j'} = \delta^{j'} W_\ell$, respectively. Since

$$L^2 = \dot{\sum_{j\in\mathbb{Z}}} W_{\ell+j} = \dot{\sum_{j'\in\mathbb{Z}}} \widetilde{W}_{\ell+j'},$$

both $\{\psi_{n;j,k}\colon\ j,k \in \mathbb{Z},\ 2^\ell \le n < 2^{\ell+1}\}$ and $\{\widetilde{\psi}_{n;j,k}\colon\ j,k \in \mathbb{Z},\ 2^\ell \le n < 2^{\ell+1}\}$ are bases of L^2. Moreover, from (2.21) and (2.22), we have

$$W_j \perp \widetilde{W}_{j'} \quad (\text{or } W_{\ell+j} \perp \widetilde{W}_{\ell+j'}), \quad j \ne j'; \quad j, j', \ell \in \mathbb{Z}. \tag{4.3}$$

Thus, (2.31) follows from (4.3) for $j \ne j'$. The assertion for $j' = j$ is a consequence of Theorem 1, namely,

$$\langle \psi_{n;j,k},\ \psi_{n';j,k'} \rangle = \langle \psi_n(\cdot - k),\ \psi_{n'}(\cdot - k') \rangle = \delta_{n,n'}\delta_{k,k'}.$$

This completes the proof of Theorem 3. □

Proof of Theorem 4. Fix any $\ell \in \mathbb{Z}_+$ and consider n with $2^\ell \le n < 2^{\ell+1}$. As seen in the proof of Theorem 3, $\{\psi_{n;\ell,k}\colon\ k \in \mathbb{Z},\ 2^\ell \le n < 2^{\ell+1}\}$ and $\{\psi_{n;\ell+1,k}\colon\ k \in \mathbb{Z},\ 2^\ell \le n < 2^{\ell+1}\}$ are bases of $W_{2\ell}$ and $W_{2\ell+1}$, respectively. Since

$$L^2 = V_0 \dotplus \dot{\sum_{\ell\in\mathbb{Z}_+}} W_\ell = U_0 \dotplus \dot{\sum_{\ell\in\mathbb{Z}_+}} W_{2\ell} + \sum_{\ell\in\mathbb{Z}_+} W_{2\ell+1},$$

it follows that the family of functions given in (2.33) is a basis of L^2. Similarly, the family given in (2.34) is also a basis of L^2. Hence, Theorem 4 follows by applying Theorems 1 and 2 as well as (4.3). □

Proof of Theorem 5. We first observe that (2.31) holds for all (n,j), $(n',j') \in \mathcal{J}$ and $k, k' \in \mathbb{Z}$. Indeed, as seen in the proof of Theorem 3, (2.31) holds for all $j' = j$, $n, n' \in \mathbb{Z}_+$ and $k, k' \in \mathbb{Z}$, no matter how $(n,j) \in \mathcal{J}$ or $(n',j') \in \mathcal{J}$ is chosen. Thus, it remains to show that if $j \ne j'$, $(n,j) \in \mathcal{J}$, $(n',j') \in \mathcal{J}$, and $k, k' \in \mathbb{Z}$, then

$$\langle \psi_{n;j,k}, \widetilde{\psi}_{n';j',k'} \rangle = 0. \tag{4.4}$$

Without loss of generality, we assume $j > j'$ and let $\ell := j - j'$, so that

$$\langle \psi_{n;j,k}, \widetilde{\psi}_{n';j',k'} \rangle = \langle \psi_{n;\ell,k}, \widetilde{\psi}_{n';0,k'} \rangle. \tag{4.5}$$

Here, it is clear that

$$\begin{aligned} &\widetilde{\psi}_{n';0,k'} \in \widetilde{U}_{n'} \\ &\psi_{n;\ell,k} \in \delta^\ell U_n = U_{2^\ell n} \dotplus \cdots \dotplus U_{2^\ell n + 2^\ell - 1}. \end{aligned} \tag{4.6}$$

Since $\{I_{n,j} := [2^j n, 2^j(n+1))\colon\ (n,j) \in \mathcal{J}\}$ forms a disjoint covering of $(0,\infty)$, the family $\{2^{-j'} I_{n,j}\colon\ (n,j) \in \mathcal{J}\}$ is also a disjoint covering of $(0,\infty)$. Thus, since $j > j'$, it follows from $(n,j) \in \mathcal{J}$, $(n',j') \in \mathcal{J}$ and $(n,j) \ne (n',j')$ that $[n', n+1)$ and $[2^\ell n, 2^\ell(n+1))$ are disjoint; or equivalently, $n' \notin [2^\ell n, 2^\ell(n+1))$, where $\ell = j - j'$. By (2.28) and (3.40), we obtain

$$\widetilde{U}_{n'} \perp \dot{\sum_{2^\ell n \le j < 2^\ell(n+1)}} U_j = \delta^\ell U_n. \tag{4.7}$$

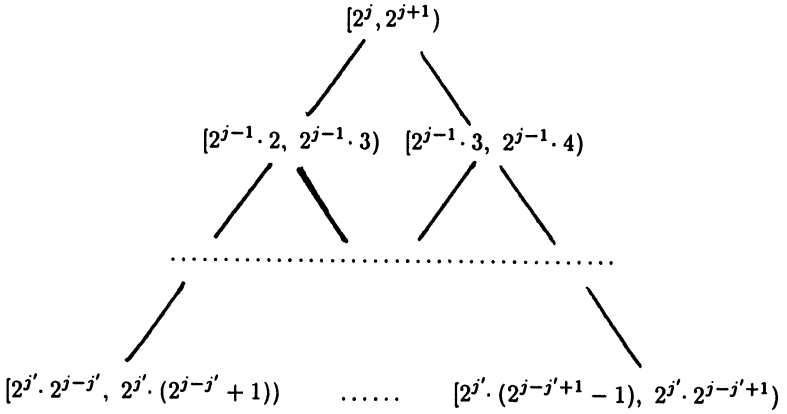

$[2^j, 2^{j+1})$
$[2^{j-1}\cdot 2,\ 2^{j-1}\cdot 3)$
$[2^{j-1}\cdot 3,\ 2^{j-1}\cdot 4)$
$[2^{j'}\cdot 2^{j-j'},\ 2^{j'}\cdot (2^{j-j'}+1))$
......
$[2^{j'}\cdot (2^{j-j'+1}-1),\ 2^{j'}\cdot 2^{j-j'+1})$

FIG. 1

Hence, (4.4) follows from (4.5), (4.6), and (4.7).

To complete the proof of Theorem 5, we have to show that $\{\psi_{n;j,k}\colon (n,j) \in \mathcal{J}, k \in \mathbb{Z}\}$ is complete in $L^2(\mathbb{R})$. Since the wavelet basis $\{\psi_{1;j,k}\colon j, k \in \mathbb{Z}\}$ is complete for $L^2(\mathbb{R})$, it is sufficient to demonstrate that each $\psi_{1;j,k}$ $(j, k \in \mathbb{Z})$ is in the L^2-closure of the linear span of $\{\psi_{n;s,t}\colon (n;s) \in \mathcal{J}, t \in \mathbb{Z}\}$. Given $j \in \mathbb{Z}$, then for any $j' \in \mathbb{Z}$, $j' \leq j$ we can decompose the interval $[2^j, 2^{j+1})$ as in the binary tree in Fig. 1.

Since $[2^j, 2^{j+1})$ is only covered by finitely many dyadic intervals $I_{n,s} = [2^s n, 2^s(n+1))$, $(n,s) \in \mathcal{J}$, we can find a sufficiently small $j' \in \mathbb{Z}$, with $j' \leq j$, such that $[2^j, 2^{j+1}) \subseteq \bigcup_{(n,s)\in\mathcal{J}_j} I_{n,s}$, where $(n,s) \in \mathcal{J}$, and $I_{n,s}$ appears in the above binary tree; and since it depends on j, we denote the set of these (n,s) by $\mathcal{J}_j$. Clearly $\mathcal{J}_j \subseteq \mathcal{J}$. For $k \in \mathbb{Z}$, we have $\psi_{1;j,k} \in \delta^j U_1 = \delta^{j'}(\delta^{j-j'} U_1)$. Thus, corresponding to each $(n,s) \in \mathcal{J}_j$, by using the formula (3.38) repeatedly, starting with $\delta^{j-j'} U_1 = \delta^{j-j'-1}(\delta U_1)$, we have

$$\psi_{1;j,k} \in \delta^{j'}(\delta^{j-j'} U_1) \subseteq \sum_{(n,s)\in\mathcal{J}_j}^{\bullet} \delta^s U_n.$$

This shows that $\psi_{1;j,k}$ is in the L^2-closure of the linear span of $\psi_{n;s,t}$ with $(n,s) \in \mathcal{J}_j \subset \mathcal{J}$ and $t \in \mathbb{Z}$; and consequently, $\{\psi_{n;j,k}\colon (n,j) \in \mathcal{J}, k \in \mathbb{Z}\}$ is complete in L^2. The same proof is valid for $\{\widetilde{\psi}_{n;j,k}\colon (n,j) \in \mathcal{J}, k \in \mathbb{Z}\}$. This completes the proof of Theorem 5. □

Remark. A formulation of Theorem 5 for the orthonormal setting (i.e., $\psi_n = \widetilde{\psi}_n$, $n \in \mathbb{Z}_+$) can be found in Coifman and Meyer [5].

Proof of Theorem 6. We first prove that, for each integer n satisfying $2^\ell \leq n < 2^{\ell+1}$, there exists some positive constant C_n, such that

$$\sum_{k\in\mathbb{Z}} |\langle f, \psi_{n;j,k}\rangle|^2 \leq C_n \sum_{k\in\mathbb{Z}} |\langle f, \psi_{1;j+\ell,k}\rangle|^2 \tag{4.8}$$

for all $j \in \mathbb{Z}$ and $f \in L^2(\mathbb{R})$. For this purpose, we will adopt the dyadic expansion for n, namely: $n = \sum_{t=1}^{\ell+1} \varepsilon_t 2^{t-1}$, where $\varepsilon_t \in \{0,1\}$ and $\varepsilon_{\ell+1} = 1$. By using the operator $\mathcal{P}_\lambda$ defined in (3.11), it follows from (3.22) that

$$\psi_{n;j,k} = 2^{-\frac{1}{2}}(\mathcal{P}_{\varepsilon_1}\{\psi_{\lfloor\frac{n}{2}\rfloor,j+1,\cdot}\})_k, \qquad k \in \mathbb{Z}, \tag{4.9}$$

or equivalently, in view of (3.11),

$$\psi_{n;j,k}(x) = 2^{-\frac{1}{2}} \sum_{m\in\mathbb{Z}} p^{\varepsilon_1}_{m-2k} \psi_{\lfloor \frac{n}{2} \rfloor; j+1,m}(x), \qquad j,k \in \mathbb{Z}. \tag{4.10}$$

Thus, we have, for any $f \in L^2$,

$$\begin{aligned} |\langle f, \psi_{n;j,k}\rangle| &\le 2^{-\frac{1}{2}} \sum_{m\in\mathbb{Z}} |p^{\varepsilon_1}_{m-2k}|^{\frac{1}{2}} \cdot |p^{\varepsilon_1}_{m-2k}|^{\frac{1}{2}} \cdot |\langle f, \psi_{\lfloor \frac{n}{2} \rfloor; j+1,m}\rangle| \\ &\le 2^{-\frac{1}{2}} \left(\sum_{m\in\mathbb{Z}} |p^{\varepsilon_1}_{m-2k}| \right)^{\frac{1}{2}} \left(\sum_{m\in\mathbb{Z}} |p^{\varepsilon_1}_{m-2k}| \cdot |\langle f, \psi_{\lfloor \frac{n}{2} \rfloor; j+1,m}\rangle|^2 \right)^{\frac{1}{2}}. \end{aligned}$$

Moreover, by putting

$$D_\lambda := \max\left\{ \sum_{k\in\mathbb{Z}} |p^{\lambda}_{2k}|, \sum_{k\in\mathbb{Z}} |p^{\lambda}_{2k+1}| \right\}, \qquad \lambda \in \{0,1\}, \tag{4.11}$$

we have

$$\sum_{m\in\mathbb{Z}} |p^{\varepsilon_1}_{m-2k}| = \sum_{m\in\mathbb{Z}} |p^{\varepsilon_1}_{m}| \le 2D_{\varepsilon_1},$$

and

$$\sum_{k\in\mathbb{Z}} |p^{\varepsilon_1}_{m-2k}| = \begin{cases} \sum\limits_{k\in\mathbb{Z}} |p^{\varepsilon_1}_{2k}| & \text{for even } m, \\ \sum\limits_{k\in\mathbb{Z}} |p^{\varepsilon_1}_{2k+1}| & \text{for odd } m, \end{cases}$$

so that

$$\begin{aligned} \sum_{k\in\mathbb{Z}} |\langle f, \psi_{n;j,k}\rangle|^2 &\le D_{\varepsilon_1} \sum_{m\in\mathbb{Z}} \left(\sum_{k\in\mathbb{Z}} |p^{\varepsilon_1}_{m-2k}| \right) |\langle f, \psi_{\lfloor \frac{n}{2} \rfloor; j+1,m}\rangle|^2 \\ &\le D^2_{\varepsilon_1} \sum_{m\in\mathbb{Z}} |\langle f, \psi_{\lfloor \frac{n}{2} \rfloor; j+1,m}\rangle|^2. \end{aligned} \tag{4.12}$$

Now, since $\lfloor \frac{n}{2} \rfloor = \sum_{t=1}^{\ell} \varepsilon_{t+1} 2^{t-1}$, $\lfloor \frac{n}{2^2} \rfloor = \sum_{t=1}^{\ell-1} \varepsilon_{t+2} 2^{t-1}$, etc., repeated applications of the inequality (4.12) to $\psi_{\lfloor \frac{n}{2} \rfloor; j+1,k}$, $\psi_{\lfloor \frac{n}{2^2} \rfloor; j+2,k}, \dots, \psi_{1;j+\ell,k}$, yield

$$\sum_{k\in\mathbb{Z}} |\langle f, \psi_{n;j,k}\rangle|^2 \le (D_{\varepsilon_1} \cdots D_{\varepsilon_\ell})^2 \sum_{k\in\mathbb{Z}} |\langle f; \psi_{1;j+\ell,k}\rangle|^2. \tag{4.13}$$

Thus, (4.8) follows from (4.13) by setting

$$C_n := (D_{\varepsilon_1} D_{\varepsilon_2} \cdots D_{\varepsilon_\ell})^2. \tag{4.14}$$

Observe that C_n is a (finite) positive constant, since it follows from (4.11) and $\{p^{\lambda}_{k}\} \in \ell^1$ that D_λ is finite, $\lambda \in \{0,1\}$. In addition, by setting $D := \max\{D_0, D_1\}$, (4.14) implies that $C_n \le D^{2\ell}$. Thus, from (4.13), we have

$$\sum_{k\in\mathbb{Z}} \sum_{2^\ell \le n < 2^{\ell+1}} |\langle f, \psi_{n;j,k}\rangle|^2 \le 2^\ell D^{2\ell} \sum_{k\in\mathbb{Z}} |\langle f, \psi_{1;j+\ell,k}\rangle|^2. \tag{4.15}$$

If $\{\psi_{1;j,k}\colon\ j,k \in \mathbb{Z}\}$ constitutes a frame of L^2, then there exists a positive constant B, such that

$$\sum_{j,k\in\mathbf{Z}} |\langle f, \psi_{1;j,k}\rangle|^2 \le B\|f\|^2.$$

Thus, combining this inequality with (4.15) gives

$$\sum_{j,k\in\mathbf{Z}}\ \sum_{2^\ell\le n<2^{\ell+1}} |\langle f, \psi_{n;j,k}\rangle|^2 \le 2^\ell D^{2\ell} B\|f\|^2. \tag{4.16}$$

Similarly, by letting

$$\widetilde{D}_\lambda := \max\left\{\sum_{k\in\mathbf{Z}} |g^\lambda_{2k}|, \sum_{k\in\mathbf{Z}} |g^\lambda_{2k+1}|\right\}, \qquad \widetilde{D} := \max\{\widetilde{D}_0, \widetilde{D}_1\} \tag{4.17}$$

and assuming that $\{\widetilde{\psi}_{1;j,k}\colon\ j,k \in \mathbb{Z}\}$ is a frame of L^2, we have

$$\sum_{j,k\in\mathbf{Z}}\ \sum_{2^\ell\le n<2^{\ell+1}} |\langle f, \widetilde{\psi}_{n;j,k}\rangle|^2 \le 2^\ell \widetilde{D}^{2\ell} \widetilde{B}\|f\|^2, \qquad f \in L^2 \tag{4.18}$$

for some positive constant $\widetilde{B}$.

To complete the proof of Theorem 6, we still have to derive the lower estimates. Now, since $L^2 = \sum^{\bullet}_{j\in\mathbf{Z}} W_{j+\ell}$, we can write, for any $f \in L^2$,

$$f = \sum_{j\in\mathbf{Z}} g_{j+\ell}, \qquad g_{j+\ell} \in W_{j+\ell}, \quad j \in \mathbb{Z}. \tag{4.19}$$

Moreover, from Theorem 3 as well as its proof, we have

$$g_{j+\ell} = \sum_{k\in\mathbf{Z}}\ \sum_{2^\ell\le n<2^{\ell+1}} \langle g_{j+\ell}, \widetilde{\psi}_{n;j,k}\rangle \psi_{n;j,k} = \sum_{k\in\mathbf{Z}}\ \sum_{2^\ell\le n<2^{\ell+1}} \langle f, \widetilde{\psi}_{n;j,k}\rangle \psi_{n;j,k}, \tag{4.20}$$

where $\langle f, \widetilde{\psi}_{n;j,k}\rangle = \langle g_{j+\ell}, \widetilde{\psi}_{n;j,k}\rangle$ follows from (4.19) and the fact that $g_{j'+\ell} \in W_{j'+\ell} \perp \widetilde{W}_{j+\ell}$ and $\widetilde{\psi}_{n;j,k} \in \widetilde{W}_{j+\ell}$ for $j' \ne j$ (see (4.3)). Thus, from (4.19) and (4.20), it follows that

$$f = \sum_{j,k\in\mathbf{Z}}\ \sum_{2^\ell\le n<2^{\ell+1}} \langle f, \widetilde{\psi}_{n;j,k}\rangle \psi_{n;j,k}, \qquad f \in L^2. \tag{4.21}$$

By the same argument, we also have

$$f = \sum_{j,k\in\mathbf{Z}}\ \sum_{2^\ell\le n<2^{\ell+1}} \langle f, \psi_{n;j,k}\rangle \widetilde{\psi}_{n;j,k}, \qquad f \in L^2. \tag{4.22}$$

Hence, by (2.31), (4.21), and (4.22), we obtain, for any $f \in L^2$,

$$\begin{aligned} \|f\|^2 = \langle f, f\rangle &= \sum_{j,k\in\mathbf{Z}}\ \sum_{2^\ell\le n<2^{\ell+1}} \langle f, \psi_{n;j,k}\rangle \overline{\langle f, \widetilde{\psi}_{n;j,k}\rangle} \\ &\le \left(\sum_{j,k\in\mathbf{Z}}\ \sum_{2^\ell\le n<2^{\ell+1}} |\langle f, \psi_{n;j,k}\rangle|^2\right)^{\frac{1}{2}} \left(\sum_{j,k\in\mathbf{Z}}\ \sum_{2^\ell\le n<2^{\ell+1}} |\langle f, \widetilde{\psi}_{n;j,k}\rangle|^2\right)^{\frac{1}{2}}. \end{aligned} \tag{4.23}$$

On the other hand, from (4.16), (4.18), and (4.23), we conclude that

$$\sum_{j,k\in\mathbb{Z}} \sum_{2^\ell \le n < 2^{\ell+1}} |\langle f, \psi_{n;j,k}\rangle|^2 \ge 2^{-\ell}\widetilde{D}^{-2\ell}\widetilde{B}^{-1}\|f\|^2, \tag{4.24}$$

and

$$\sum_{j,k\in\mathbb{Z}} \sum_{2^\ell \le n < 2^{\ell+1}} |\langle f, \widetilde{\psi}_{n;j,k}\rangle|^2 \ge 2^{-\ell}D^{-2\ell}B^{-1}\|f\|^2. \tag{4.25}$$

A combination of the inequalities (4.16), (4.18), (4.24), and (4.25) completes the proof of Theorem 6. □

5. Algorithms for wavelets packets. In this section, we will give two change-of-bases algorithms using wavelet packets.

Given a function $f \in L^2$, we approximate f on the scale 2^{-N} by some

$$f_N(x) := \sum_{j\in\mathbb{Z}} \bar{c}_j \psi_0(2^N x - j) = \langle\{\psi_0(2^N x - \cdot)\}, \mathbf{c}\rangle_{\ell^2} \tag{5.1}$$

in V_N, where $\mathbf{c} = \{c_j\}_{j\in\mathbb{Z}} \in \ell^2$. Thus, from the formula

$$\psi_k(2x - j) = \tfrac{1}{2}(\mathcal{G}_0^*\{\psi_{2k}(x - \cdot)\})_j + \tfrac{1}{2}(\mathcal{G}_1^*\{\psi_{2k+1}(x - \cdot)\})_j \tag{5.2}$$

(see (3.24)), where $j \in \mathbb{Z}$ and $k \in \mathbb{Z}_+$, we obtain

$$\begin{aligned} f_N(x) &= \frac{1}{2}\langle \mathcal{G}_0^*\{\psi_0(2^{N-1}x - \cdot)\}, \mathbf{c}\rangle_{\ell^2} + \frac{1}{2}\langle \mathcal{G}_1^*\{\psi_1(2^{N-1}x - \cdot)\}, \mathbf{c}\rangle_{\ell^2} \\ &= \frac{1}{2}\langle \{\psi_0(2^{N-1}x - \cdot)\}, \mathcal{G}_0\mathbf{c}\rangle_{\ell^2} + \frac{1}{2}\langle \{\psi_1(2^{N-1}x - \cdot)\}, \mathcal{G}_1\mathbf{c}\rangle_{\ell^2} \\ &= \frac{1}{2^2}\langle \mathcal{G}_0^*\{\psi_0(2^{N-2}x - \cdot)\}, \mathcal{G}_0\mathbf{c}\rangle_{\ell^2} + \frac{1}{2^2}\langle \mathcal{G}_1^*\{\psi_1(2^{N-2}x - \cdot)\}, \mathcal{G}_0\mathbf{c}\rangle_{\ell^2} \\ &\quad + \frac{1}{2^2}\langle \mathcal{G}_0^*\{\psi_2(2^{N-2}x - \cdot)\}, \mathcal{G}_1\mathbf{c}\rangle_{\ell^2} + \frac{1}{2^2}\langle \mathcal{G}_1^*\{\psi_3(2^{N-2}x - \cdot)\}, \mathcal{G}_1\mathbf{c}\rangle_{\ell^2} \\ &= \frac{1}{2^2}\langle \{\psi_0(2^{N-2}x - \cdot)\}, \mathcal{G}_0^2\mathbf{c}\rangle_{\ell^2} + \frac{1}{2^2}\langle \{\psi_1(2^{N-2}x - \cdot)\}, \mathcal{G}_1\mathcal{G}_0\mathbf{c}\rangle_{\ell^2} \\ &\quad + \frac{1}{2^2}\langle \{\psi_2(2^{N-2}x - \cdot)\}, \mathcal{G}_0\mathcal{G}_1\mathbf{c}\rangle_{\ell^2} + \frac{1}{2^2}\langle \{\psi_3(2^{N-2}x - \cdot)\}, \mathcal{G}_1^2\mathbf{c}\rangle_{\ell^2} \\ &= \dots, \end{aligned} \tag{5.3}$$

where the operators $\mathcal{G}_0$ and $\mathcal{G}_1$ are defined in (3.12). Finally we have, for $\ell = 1, 2, \dots, M$,

$$f_N(x) = \frac{1}{2^\ell}\sum_{k=0}^{2^\ell - 1} \langle \{\psi_k(2^{N-\ell}x - \cdot)\}, \mathcal{G}_{k,\ell}\mathbf{c}\rangle_{\ell^2}, \tag{5.4}$$

where for $k = \sum_{j=1}^{\ell} \varepsilon_j 2^{j-1}$, $\varepsilon_j \in \{0, 1\}$,

$$\mathcal{G}_{k,\ell} := \mathcal{G}_{\varepsilon_1}\mathcal{G}_{\varepsilon_2}\cdots\mathcal{G}_{\varepsilon_\ell}. \tag{5.5}$$

That is, f_N has the unique decomposition

$$f_N = \sum_{k=0}^{2^\ell-1} g_{k,\ell} \in \delta^{N-\ell} \sum_{0\le k<2^\ell}^{\bullet} U_k = V_N, \tag{5.6}$$

where $g_{k,l} := \frac{1}{2^\ell}\langle\{\psi_k(2^{N-\ell}x-\cdot)\},\, \mathcal{G}_{k,\ell}\mathbf{c}\rangle_{\ell^2} \in \delta^{N-\ell}U_k$. This decomposition algorithm can be expressed recursively as follows:

$$\mathbf{c}^{k,\ell} = \tfrac{1}{2}\mathcal{G}_{\lambda_k}\mathbf{c}^{\lfloor k/2\rfloor,\ell-1}, \quad \ell = 1,\dots,M; \quad k = 0,\dots,2^\ell-1, \tag{5.7}$$

where $\mathbf{c}^{0,0} := \mathbf{c}$, $\lambda_k \in \{0,1\}$, satisfies $k = 2\lfloor\frac{k}{2}\rfloor + \lambda_k$; i.e., $\lambda_k = 0$ for even k, and $\lambda_k = 1$ for odd k. The procedure can be described by the binary tree as shown in Fig. 2. Here,

$$f_N(x) = \langle\{\psi_0(2^Nx-\cdot)\},\mathbf{c}\rangle_{\ell^2} = \sum_{k=0}^{2^\ell-1}\langle\{\psi_k(2^{N-\ell}x-\cdot)\},\mathbf{c}^{k,\ell}\rangle_{\ell^2}, \tag{5.8}$$

where $\ell = 1,2,\dots,M$.

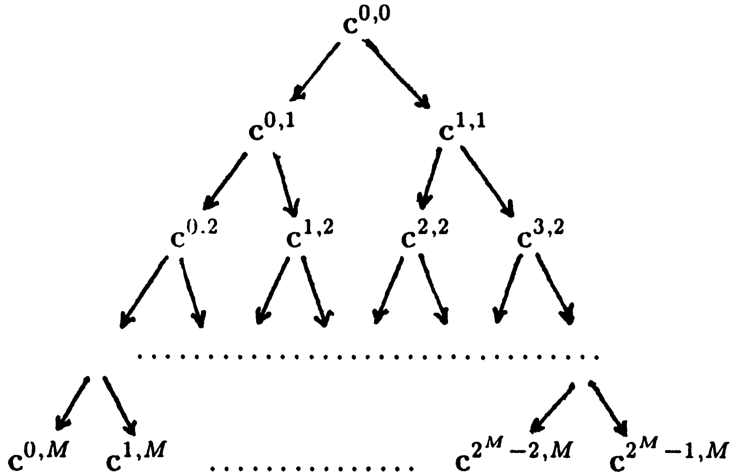


FIG. 2

On the other hand, for the reconstruction algorithm it follows from the representation (5.8) and the formula

$$\psi_k(x-j) = (\mathcal{P}_{\lambda_k}\{\psi_{\lfloor\frac{k}{2}\rfloor}(2x-\cdot)\})_j, \qquad j\in\mathbb{Z}, \tag{5.9}$$

where $\lambda_k\in\{0,1\}$ satisfies $k = 2\lfloor\frac{k}{2}\rfloor + \lambda_k$, that

$$\begin{aligned} f_N(x) &= \sum_{k=0}^{2^\ell-1}\langle\mathcal{P}_{\lambda_k}\{\psi_{\lfloor\frac{k}{2}\rfloor}(2^{N-\ell+1}x-\cdot)\},\mathbf{c}^{k,\ell}\rangle_{\ell^2} \\ &= \sum_{k=0}^{2^\ell-1}\langle\{\psi_{\lfloor\frac{k}{2}\rfloor}(2^{N-\ell+1}x-\cdot)\},\mathcal{P}^*_{\lambda_k}\mathbf{c}^{k,\ell}\rangle_{\ell^2} \\ &= \sum_{k=0}^{2^{\ell-1}-1}\langle\{\psi_k(2^{N-\ell+1}x-\cdot)\},\mathbf{c}^{k,\ell-1}\rangle_{\ell^2}, \end{aligned} \tag{5.10}$$

where

$$\mathbf{c}^{k,\ell-1} := \mathcal{P}_0^*\mathbf{c}^{2k,\ell} + \mathcal{P}_1^*\mathbf{c}^{2k+1,\ell}, \tag{5.11}$$

$\ell = M, M-1, \dots, 1$; $k = 0, 1, \dots, 2^{\ell-1}-1$, and the operators $\mathcal{P}_0^*$ and $\mathcal{P}_1^*$ are given in (3.13). Finally, for $\ell = 1$, we obtain

$$f_N(x) = \langle\{\psi_0(2^N x - \cdot)\}, \mathbf{c}^{0,0}\rangle_{\ell^2} = \langle\{\psi_0(2^N x - \cdot)\}, \mathbf{c}\rangle_{\ell^2}.$$

The procedure for the reconstruction algorithm can be described by the binary tree as shown in Fig. 3.

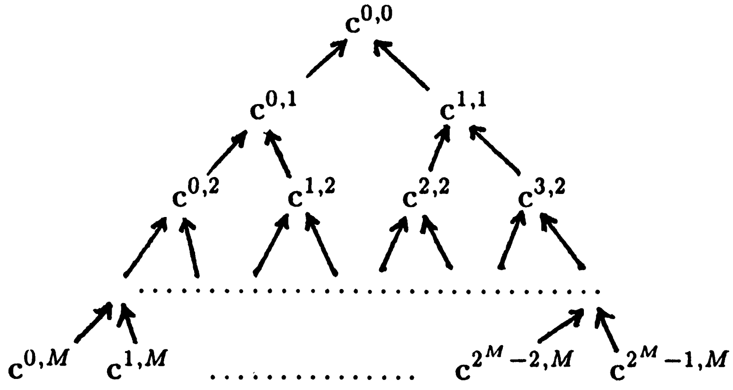


FIG. 3.

Both of the above decomposition and reconstruction algorithms are given relative to the wavelet packets $\{\psi_k\}$ and the decomposition

$$V_N = \delta^{N-\ell}\delta^{\ell}V_0 = \delta^{N-\ell}\overset{\bullet}{\sum_{0\le k<2^\ell}} U_k.$$

The computational complexity of these algorithms is $O(n\log n)$, where $n = 2^M$ is the number of operations for precomputing all the ψ_k, $k = 0, \dots, 2^M-1$. In the following we give another algorithms corresponding to Theorem 3.

Let $\ell \in \mathbb{Z}_+$ and $M \in \mathbb{Z}_+$ be fixed. As before, to a given $f \in L^2$, let f_N approximate f from V_N. Then since

$$V_N = W_{N-1}\dot{+}V_{N-1} = \cdots = W_{N-1}\dot{+}W_{N-2}\dot{+}\cdots\dot{+}W_{N-M}\dot{+}V_{N-M},$$

f_N has a unique decomposition

$$f_N = g_{N-1} + g_{N-2} + \cdots + g_{N-M} + f_{N-M}, \tag{5.12}$$

where $f_{N-M} \in V_{N-M}$ and $g_j \in W_j$, $j = N-M, \dots, N-1$. Write

$$f_j(x) = \sum_{k\in\mathbf{Z}} \overline{c_k^j}\psi_0(2^j x - k) = \langle\{\psi_0(2^j x - \cdot)\}, \mathbf{c}^j\rangle_{\ell^2},$$

and

$$g_j(x) = \sum_{k\in\mathbf{Z}} \overline{d_k^j}\psi_1(2^j x - k) = \langle\{\psi_1(2^j x - \cdot)\}, \mathbf{d}^j\rangle_{\ell^2},$$

where $\mathbf{c}^j = \{c_k^j\}_{k\in\mathbf{Z}} \in \ell^2$ and $\mathbf{d}^j = (d_k^j)_{k\in\mathbf{Z}} \in \ell^2$. From the decomposition formula (5.2), we have, for $k = 0$,

$$\begin{aligned} f_j(x) &= \left\langle \tfrac{1}{2}\mathcal{G}_0^*\{\psi_0(2^{j-1}x - \cdot)\}, \mathbf{c}^j \right\rangle_{\ell^2} + \left\langle \tfrac{1}{2}\mathcal{G}_1^*\{\psi_1(2^{j-1}x - \cdot)\}, \mathbf{c}^j \right\rangle_{\ell^2} \\ &= \left\langle \{\psi_0(2^{j-1}x - \cdot)\}, \tfrac{1}{2}\mathcal{G}_0\mathbf{c}^j \right\rangle_{\ell^2} + \left\langle \{\psi_1(2^{j-1}x - \cdot)\}, \tfrac{1}{2}\mathcal{G}_1\mathbf{c}^j \right\rangle_{\ell^2} \\ &= f_{j-1}(x) + g_{j-1}(x). \end{aligned} \tag{5.13}$$

Thus, it follows that

$$\mathbf{c}^{j-1} = \tfrac{1}{2}\mathcal{G}_0\mathbf{c}^j, \quad \mathbf{d}^{j-1} = \tfrac{1}{2}\mathcal{G}_1\mathbf{c}^j, \qquad j = N, N-1, \ldots, N-M+1. \tag{5.14}$$

This procedure is described by the graph in Fig. 4.

$$\begin{tikzcd} & \mathbf{d}^{N-1} & \mathbf{d}^{N-2} & & \mathbf{d}^{N-M} \\ \mathbf{c}^N \arrow[r] \arrow[ur] & \mathbf{c}^{N-1} \arrow[r] \arrow[ur] & \mathbf{c}^{N-2} \arrow[r] \arrow[ur] & \cdots \arrow[r] \arrow[ur] & \mathbf{c}^{N-M} \end{tikzcd}$$

FIG. 4

Moreover, since

$$g_j \in W_j = \delta^j W_1 = \delta^{j-\ell}\delta^\ell W_1 = \delta^{j-\ell} \dot{\sum_{2^\ell \le k < 2^{\ell+1}}} U_k,$$

then by applying the decomposition formula (5.2) again for $k = 1, 2, \ldots, 2^{\ell-1} - 1$, we have

$$\begin{aligned} g_j(x) &= \langle \{\psi_1(2^j - x\cdot)\}, \mathbf{d}^j \rangle_{\ell^2} \\ &= \frac{1}{2}\langle \mathcal{G}_0^*\{\psi_2(2^{j-1}x - \cdot)\}, \mathbf{d}^j \rangle_{\ell^2} + \frac{1}{2}\langle \mathcal{G}_1^*\{\psi_3(2^{j-1}x - \cdot)\}, \mathbf{d}^j \rangle_{\ell^2} \\ &= \frac{1}{2}\langle \{\psi_2(2^{j-1}x - \cdot)\}, \mathcal{G}_0\mathbf{d}^j \rangle_{\ell^2} + \frac{1}{2}\langle \{\psi_3(2^{j-1}x - \cdot)\}, \mathcal{G}_1\mathbf{d}^j \rangle_{\ell^2} \\ &= \cdots \\ &= \frac{1}{2^\ell} \sum_{k=2^\ell}^{2^{\ell+1}-1} \langle \{\psi_k(2^{j-\ell}x - \cdot)\}, \mathcal{G}_{k,\ell}\mathbf{d}^j \rangle_{\ell^2}, \end{aligned} \tag{5.15}$$

where the operator $\mathcal{G}_{k,\ell}$ is defined on ℓ^2 by (5.5). Writing

$$\mathbf{a}^{j,k,m} := \mathcal{G}_{k,m}\mathbf{d}^j, \qquad \mathbf{a}^{j,1,0} := \mathbf{d}^j, \tag{5.16}$$

we have, similar to (5.7), that

$$\mathbf{a}^{j,k,m} = \tfrac{1}{2}\mathcal{G}_{\lambda_k}\mathbf{a}^{j,\lfloor \frac{k}{2} \rfloor, m-1}, \tag{5.17}$$

where $m = 1, \ldots, \ell$, $k = 2^m, \ldots, 2^{m+1} - 1$, and $\lambda_k \in \{0, 1\}$, satisfying $k = 2\lfloor \frac{k}{2} \rfloor + \lambda_k$. Thus, we can finally conclude that

(5.18)

$$\begin{aligned} f_N(x) &= f_{N-M}(x) + \sum_{j=N-M}^{N-1} g_j(x) \\ &= \langle \{\psi_0(2^{N-M}x - \cdot)\}, \mathbf{c}^{N-M} \rangle_{\ell^2} + \sum_{j=N-M}^{N-1} \langle \{\psi_1(2^j x - \cdot)\}, \mathbf{d}^j \rangle_{\ell^2} \\ &= \langle \{\psi_0(2^{N-M}x - \cdot)\}, \mathbf{c}^{N-M} \rangle_{\ell^2} + \sum_{j=N-M}^{N-1} \sum_{2^\ell \le k < 2^{\ell+1}} \langle \{\psi_k(2^{j-\ell}x - \cdot)\}, \mathbf{a}^{j,k,\ell} \rangle_{\ell^2}, \end{aligned}$$

where $\mathbf{c}^{N-M}, \mathbf{d}^j$ and $\mathbf{a}^{j,k,\ell}$, $N-M \le j \le N-1$, $2^\ell \le k < 2^{\ell+1}$, are given by (5.14) and (5.17), respectively. This is the decomposition algorithm corresponding to Theorem 3.

On the other hand, if we start with

$$g_j(x) = \sum_{2^m \le k < 2^{m+1}} \langle \{\psi_k(2^{j-m}x - \cdot)\}, \mathbf{a}^{j,k,m} \rangle_{\ell^2}, \tag{5.19}$$

$m = \ell, \ell-1, \ldots, 1$, for some $\mathbf{a}^{j,k,m} \in \ell^2$; then by using (5.9) again, we obtain

$$\begin{aligned} g_j(x) &= \sum_{2^m \le k < 2^{m+1}} \langle \mathcal{P}_{\lambda_k}\{\psi_{\lfloor \frac{k}{2} \rfloor}(2^{j-m+1}x - \cdot)\}, \mathbf{a}^{j,k,m} \rangle_{\ell^2} \\ &= \sum_{2^m \le k < 2^{m+1}} \langle \{\psi_{\lfloor \frac{k}{2} \rfloor}(2^{j-m+1}x - \cdot)\}, \mathcal{P}^*_{\lambda_k}\mathbf{a}^{j,k,m} \rangle_{\ell^2} \\ &= \sum_{2^{m-1} \le k < 2^m} \langle \{\psi_k(2^{j-m+1}x - \cdot)\}, \mathbf{a}^{j,k,m-1} \rangle_{\ell^2}, \end{aligned}$$

where

$$\mathbf{a}^{j,k,m-1} := \mathcal{P}_0^*\mathbf{a}^{j,2k,m} + \mathcal{P}_1^*\mathbf{a}^{j,2k+1,m}, \tag{5.20}$$

$m = \ell, \ell-1, \ldots, 1$; $k = 2^{m-1}, \ldots, 2^m - 1$. Thus we obtain, for $N - M \le j \le N-1$,

$$g_j(x) = \langle \{\psi_1(2^j x - \cdot)\}, \mathbf{a}^{j,1,0} \rangle_{\ell^2} = \langle \{\psi_1(2^j x - \cdot)\}, \mathbf{d}^j \rangle_{\ell^2}. \tag{5.21}$$

With the above sequences $\mathbf{d}^j = \mathbf{a}^{j,1,0}$ and some $\mathbf{c}^{N-M} \in \ell^2$, where $f_{N-M}(x) = \langle \{\psi_0(2^{N-M}x - \cdot)\}, \mathbf{c}^{N-M} \rangle_{\ell^2}$, we can reconstruct $f_N(x) = \langle \{\psi_0(2^N x - \cdot)\}, \mathbf{c}^N \rangle_{\ell^2}$ by applying the formula

$$\mathbf{c}^j = \mathcal{P}_0^*\mathbf{c}^{j-1} + \mathcal{P}_1^*\mathbf{d}^{j-1}, \qquad j = N - M + 1, \ldots, N. \tag{5.22}$$

To see this, we simply observe that

$$\begin{aligned} \langle \{\psi_0(2^j x - \cdot)\}, \mathbf{c}^j \rangle_{\ell^2} &= f_j(x) = f_{j-1}(x) + g_{j-1}(x) \\ &= \langle \{\psi_0(2^{j-1}x - \cdot)\}, \mathbf{c}^{j-1} \rangle_{\ell^2} + \langle \{\psi_1(2^{j-1}x - \cdot)\}, \mathbf{d}^{j-1} \rangle_{\ell^2} \\ &= \langle \mathcal{P}_0\{\psi_0(2^j x - \cdot)\}, \mathbf{c}^{j-1} \rangle_{\ell^2} + \langle \mathcal{P}_1\{\psi_0(2^j x - \cdot)\}, \mathbf{d}^{j-1} \rangle_{\ell^2} \\ &= \langle \{\psi_0(2^j x - \cdot)\}, \mathcal{P}_0^*\mathbf{c}^{j-1} \rangle_{\ell^2} + \langle \{\psi_0(2^j x - \cdot)\}, \mathcal{P}_1^*\mathbf{d}^{j-1} \rangle_{\ell^2} \\ &= \langle \{\psi_0(2^j x - \cdot)\}, \mathcal{P}_0^*\mathbf{c}^{j-1} + \mathcal{P}_1^*\mathbf{d}^{j-1} \rangle_{\ell^2}. \end{aligned}$$

This procedure is described in Fig. 5.

$$\begin{tikzcd}
\mathbf{d}^{N-M} \arrow[dr] & \mathbf{d}^{N-M+1} \arrow[dr] & {} \arrow[dr] & \mathbf{d}^{N-1} \arrow[dr] & \\
\mathbf{c}^{N-M} \arrow[r] & \mathbf{c}^{N-M+1} \arrow[r] & \cdots \arrow[r] & \mathbf{c}^{N-1} \arrow[r] & \mathbf{c}^{N}
\end{tikzcd}$$

FIG. 5

Acknowledgment. We wish to express our appreciation to Albert Cohen for his valuable comments on the first version of this manuscript and for pointing out a very interesting negative result on stability [3]. We would also like to acknowledge the very useful comments from the referees that helped improve the final draft of the manuscript.

REFERENCES

[1] C. K. CHUI, *An Introduction to Wavelets*, Academic Press, Boston, MA, 1992.

[2] C. K. CHUI AND J. Z. WANG, *On compactly supported spline wavelets and a duality principle*, Trans. Amer. Math. Soc., 330 (1992), pp. 903–915.

[3] A. COHEN AND I. DAUBECHIES, private communication, January 1992.

[4] A. COHEN, I. DAUBECHIES, AND J. C. FEAUVEAU, *Bi-orthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., to appear.

[5] R. R. COIFMAN AND Y. MEYER, *Orthonormal wave packet bases*, preprint.

[6] R. R. COIFMAN, Y. MEYER, S. QUAKE, AND M. V. WICKERHAUSER, *Signal processing and compression with wave packets*, in Proceedings of the Conference on Wavelets, Marseilles, 1989, to appear.

[7] I. DAUBECHIES, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[8] R. J. DUFFIN AND A. C. SCHAEFFER, *A class of nonharmonic Fourier series*, Trans. Amer. Math. Soc., 72 (1952), pp. 341–366.

[9] S. MALLAT, *Multiresolution representation and wavelets*, Ph.D. thesis, Dept. of Elec. Engineering, Univ. of Pennsylvania, Philadelphia, PA, 1988.

[10] Y. MEYER, *Principe d'incertitude, bases Hilbertiennes et algèbres d'opérateurs*, Sém. Bourbaki, 662, 1985–1986.

[11] M. V. WICKERHAUSER, *Acoustic signal compression with wavelet packets*, in Wavelets: A Tutorial in Theory and Applications, C. K. Chui, ed., Academic Press, Boston, MA, 1992, pp. 679–700.

SIAM J. MATH. ANAL.
Vol. 24, No. 3, pp. 739–755, May 1993

CONTINUOUS WAVELET DECOMPOSITIONS, MULTIRESOLUTION, AND CONTRAST ANALYSIS*

M. DUVAL-DESTIN†, M. A. MUSCHIETTI‡, AND B. TORRESANI§

Abstract. A continuous version of multiresolution analysis is described, starting from usual continuous wavelet decompositions. Scale discretization leads to decompositions into functions of arbitrary bandwidth, satisfying QMF-like conditions. Finally, a nonlinear multiresolution scheme is described, providing multiplicative reconstruction formulas.

Key words. wavelets, multiresolution analysis, contrast analysis

AMS subject classifications. 41A58, 42C15, 44A05, 44A35

Introduction. Wavelet analysis of a function $f \in L^2(I\!R)$ basically consists in the decomposition of f as a sum of wavelets $\psi_{(b,a)}(x) = \frac{1}{a}\,\psi\left(\frac{x-b}{a}\right)$, dilated and translated copies of the mother wavelet ψ, an integrable function with vanishing integral. The coefficients of the decomposition are nothing but the scalar products $\langle f, \psi_{(b,a)}\rangle$.

Continuous wavelet decompositions have been introduced (or reintroduced, since similar tools have been used by mathematicians for a long time to study certain classes of operators [Cal]) by Grossmann and Morlet [Gr-Mo], and many applications have been developed (see [Co-Gr-Tc] and the references therein), in particular in a signal analysis context [Gr-KM-Mo], [De]. The discretization of the continuous formulas was studied later [Da1], and handled by the introducion of the notion of frames of wavelets (see also [He-Wa]). Continuous wavelets and frames of wavelets were used later on for many problems of applied mathematics and physics, and in particular to model physiology of vision, following the program of Marr [Mar] (see also [Fr-Mo], [DD]). In particular, they allowed one of the authors to introduce the notion of scale-space contrast [DD], and to exhibit multiplicative reconstruction formulas from the contrast functions.

The discovery of orthonormal bases of wavelets by Stromberg [Str] and Meyer and his collaborators (see [Me] and the references therein) opened a new door in the understanding of the theory, making, in particular, the connection with subband coding, familiar to electrical engineers and image processors [Da2]. This, moreover, led to the construction of a large family (a library in the author's terminology) of orthonormal bases, called wavelet packets bases, providing adaptive decompositions of functions directly implementable in fast (pyramidal) algorithms [Co-M-Q-W]. A continuous analogue of wavelet packets was proposed in [To1], [To2], in which the pyramidal algorithmic structure is unfortunately lost. The key point of the construction of wavelet packets is that they allow us to control the relative bandwidth of the analyzing functions in the Fourier space (the so-called surtension factor), as a function of frequency.

*Received by the editors August 5, 1991; accepted for publication (in revised form) August 31, 1992. This work was partially supported by the Groupement De Recherches "Ondelettes," Centre National de la Recherche Scientifique.

†Direction des Armements Terrestres, Etablissement Technique d'Angers, BP 4107, 49041 Angers Cedex, France.

‡Departamento de Matemáticas, Universidad de La Plata, C.C. 172, 1900 La Plata, Argentina.

§Centre de Physique Théorique, Centre National de la Recherche Scientifique-Luminy, Case 907, 13288 Marseille Cedex 09, France.

We present here a general construction, allowing the construction of wavelet packets starting from the usual continuous wavelet analysis. By wavelet packets we mean families of functions generated from a single one by simple transformations and whose relative bandwidth is nonconstant (contrary to usual wavelets) and can be matched to a given analyzed function. Let us stress here that our construction is *not* a continuous version of that of [Co-M-Q-W], and is actually quite different. Starting from an admissible analyzing wavelet ψ, it is well known that one can associate with it a scaling function ϕ, and then mimic the multiresolution construction. This is briefly described in §1. In a similar way, the continuous decomposition over scales can be replaced by a discrete one, as shown in §2. The corresponding sequence of scale parameters can be chosen arbitrarily, and the functions appearing in the decomposition (the wavelet packets) must be matched to that sequence. As a particular case, a geometric sequence yields usual Littlewood–Paley decompositions. In §3 we continue developing the analogy between continuous wavelet analysis and multiresolution analysis. In particular, the scaling function can be expressed as a *continuous* infinite product of dilated copies of a low-pass filter, denoted by μ_0 (let us recall that a continuous product, or multiplicative integral $\prod_a^b (d\mu(x))\, f(x)$ is defined to equal $\exp[\int_a^b \ln f(x)\, d\mu(x)]$ whenever such an expression makes sense [Gui]). Then using the wavelet packets construction, a continuous finite product of such μ_0 filters yields new low-pass filters, denoted by m_0, which makes the connection with the quadrature mirror filters (QMFs) appearing in multiresolution analysis. Section 4 is devoted to the construction of contrast analysis. We define the infinitesimal contrast function $c_a(x)$ of $f(x)$ as (roughly speaking) the quotient of the details of $f(x)$ at scale a by the approximation of $f(x)$ at scale a. Then under some analyticity assumptions on $|\hat{\phi}|^2$, we show that such contrast functions provide a multiplicative decomposition of $f(x)$ over scales. Otherwise stated, $f(x)$ can be continuously factorized into exponentials of contrast functions. Such continuous product formulas can be discretized in the same way as in §3, yielding very simple factorization formulas, still interpretable in terms of the integrated contrast functions introduced in [DD]. This procedure is closely connected to models of human vision [DD], [Me-DD-Ge] in which the emphasis is put on multiresolution image processing and logarithmic light sensitivity.

1. Infinitesimal multiresolution analysis of $L^2(\mathbb{R})$.

1.1. Bilinear analysis. Let us start from standard notions of continuous wavelet analysis. (Throughout this paper, our conventions for Hermitian product and Fourier transform in $L^2(\mathbb{R})$ are the following ones: $\langle f, g\rangle = \int f(x)g(x)^*\, dx$, where the star denotes complex conjugation and $\hat{f}(\xi) = \langle f, \epsilon_\xi\rangle$, where $\epsilon_\xi(x) = \exp(i\xi x)$.) We will focus on the analysis of $L^2(\mathbb{R})$, and sometimes describe in a few words the corresponding results in the $H^2(\mathbb{R})$ context (we will denote here by $H^2(\mathbb{R}) = \{f \in L^2(\mathbb{R}), \hat{f}(\xi) = 0 \text{ for all } \xi \leq 0\}$ the complex Hardy space).

Generically, a *wavelet* (or mother wavelet) is a function $\psi \in L^1(\mathbb{R})$ such that the following admissibility condition holds:

$$c_\psi = \int_0^\infty \left|\hat{\psi}(u)\right|^2 \frac{du}{u} = \int_0^\infty \left|\hat{\psi}(-u)\right|^2 \frac{du}{u} = 1. \tag{1.1}$$

If $\hat{\psi}$ is differentiable, (1.1) basically means that $\hat{\psi}(0) = 0$, otherwise stated as

$$\int_{-\infty}^{\infty} \psi(x)\, dx = 0.$$

Such a mother wavelet provides the following analysis of $L^2(\mathbb{R})$: for any $(b,a) \in \mathbb{R} \times \mathbb{R}_+^*$, we introduce the wavelet

$$\psi_{(b,a)}(x) = \frac{1}{a}\,\psi\left(\frac{x-b}{a}\right), \tag{1.2}$$

and we have the following.

THEOREM 1. *Let ψ be a mother wavelet. Then any $f \in L^2(\mathbb{R})$ decomposes as follows*:

$$f = \int_{\mathbb{R}\times\mathbb{R}_+^*} \langle f, \psi_{(b,a)}\rangle\, \psi_{(b,a)}\, \frac{db\,da}{a} \tag{1.3}$$

strongly in $L^2(\mathbb{R})$.

Proof. The proof follows from standard arguments, and we sketch it for completeness. Let

$$T_f(b,a) = \langle f, \psi_{(b,a)}\rangle.$$

Then $T_f \in L^2(\mathbb{R}, db)$ by Young's convolution inequality. Define

$$d_a(x) = \int_{\mathbb{R}} T_f(b,a)\, \psi_{(b,a)}(x)\, db. \tag{1.4}$$

Again, Young's inequality ensures that $d_a \in L^2(\mathbb{R})$ for any $a \in \mathbb{R}_+^*$. Moreover, setting

$$s_\varepsilon^\rho(x) = \int_\varepsilon^\rho d_a(x)\, \frac{da}{a}\ , \tag{1.5}$$

we have that $s_\varepsilon^\rho \in L^2(\mathbb{R})$ and $|\hat{s_\varepsilon^\rho}(\xi)| < |\hat{f}(\xi)|$ almost everywhere, so that the Lebesgue dominated convergence theorem implies that

$$\lim_{\varepsilon\to 0,\rho\to\infty} ||f - s_\varepsilon^\rho||_2 = 0, \tag{1.6}$$

yielding the theorem.

Theorem 1 has been known for a long time by mathematicians as Calderón's identity [Cal], [Fr-Ja-We]. It was rediscovered more recently in a signal analysis context by Grossmann and Morlet [Gr-Mo], and interpreted as follows: $T_f \in L^2(\mathbb{R} \times \mathbb{R}_+^*)$ is called the wavelet transform of f with respect to the analyzing wavelet ψ. If ψ is sufficiently well localized in time and frequency (i.e., both ψ and $\hat{\psi}$ have sufficient decay at infinity), T_f gives information on the time-frequency localization of f. Conversely, (1.3) states that the wavelet transform is invertible on its range, allowing the reconstruction of the analyzed function from its wavelet transform.

If we restrict to the Hardy space $H^2(\mathbb{R})$, a weaker admissibility condition (concerning only the positive frequency part of ψ) is sufficient. Simply assuming that

$$c_\psi = \int_0^\infty \left|\hat{\psi}(u)\right|^2 \frac{du}{u} = 1,$$

Theorem 1 holds for any $f \in H^2(\mathbb{R})$.

We will need in the sequel a somewhat stronger assumption. We will call ψ an *infinitesimal wavelet* if ψ is a *real-valued* wavelet, belonging to the atomic Hardy space $H^1_a(I\!R)$. $H^1_a(I\!R)$ is the space of the functions of $L^1(I\!R)$ such that their Hilbert transform (we recall that the Hilbert transform $H.f$ of a function f can be defined in the Fourier space as $\widehat{H}.f(\xi) = -i\,\mathrm{sgn}(\xi)\hat{f}(\xi))$ is also in $L^1(I\!R)$ (see [Co-We] for a detailed account of the theory of real Hardy spaces).

Let then $\psi(x)$ be an infinitesimal wavelet, and let $q(x)$ be its autocorrelation function:

$$q(x) = \int_{-\infty}^{\infty} \psi(y)\psi(x+y)\, dy. \tag{1.7}$$

Clearly $q \in H^1_a(I\!R)$, and we easily check that

$$\int_0^{\infty} q(x)\, dx = \int_{-\infty}^{0} q(x)\, dx = 0. \tag{1.8}$$

Let

$$\begin{aligned} p(x) &= \frac{1}{x} \int_0^x q(y)\, dy, \qquad x > 0 \\ &= -\frac{1}{x} \int_x^0 q(y)\, dy, \qquad x < 0 \end{aligned} \tag{1.9}$$

denote the mean function of $q(x)$. We also have

$$\hat{p}(\xi) = \int_1^{\infty} |\hat{\psi}(t\xi)|^2 \frac{dt}{t}\ .$$

We will also denote $p_a(x) = \frac{1}{a} p\left(\frac{x}{a}\right)$.

LEMMA. $p \in L^1(I\!R)$.

Proof. Let q^+ (respectively, q^-) denote the restriction of q to the positive (respectively, negative) real axis. Since $q^+ \in H^1_a(I\!R)$, $q^+(x)$ admits an atomic decomposition with $(1,\infty)$-atoms,

$$q^+(x) = \sum_n \lambda_n^+ a_n^+(x), \tag{1.10}$$

where

$$\sum_n |\lambda_n^+| < \infty, \tag{1.11}$$

and the $(1,\infty)$-atoms $a_n^+(x)$ are compactly supported L^1-functions with support in an interval I_n^+, such that

$$\int a_n^+(x)\, dx = 0$$

and

$$||a_n^+||_\infty \leq \frac{1}{|I_n^+|}\ .$$

Moreover, such a decomposition is not unique, and (because of (1.8)) can be chosen in such a way that $I_n^+ \subset I\!R^+$. Indeed, if $q^+(x) = \sum_n \lambda_n^+ a_n^+(x)$ is an atomic decomposition of $q^+(x)$ with $(1,\infty)$-atoms $a_n^+(x)$, then it also admits an atomic decomposition $q^+(x) = \sum_n \lambda_n^+ b_n^+(x)$, where $b_n^+(x) = [a_n^+(x) + a_n^+(-x)]\chi^+(x)$ are also $(1,\infty)$-atoms, supported on the positive real axis (here $\chi^+(x)$ is the Heaviside step function). A similar property obviously holds for q^-. We then assume from now on that $\mathrm{Supp}(a_n^\pm) \subset I\!R^\pm$. Denote $p_n^+(x) = \frac{1}{x}\int_0^x a_n^+ \, dy(y)$ and $p_n^-(x) = -\frac{1}{x}\int_x^0 a_n^-(y)\, dy$. Then

$$\text{(1.12)} \qquad ||p||_1 \le \sum_n (|\lambda_n^+|\, ||p_n^+||_1 + |\lambda_n^-|\, ||p_n^-||_1),$$

so that $||p||_1 < \infty$ if the $||p_n^\pm||_1$ are uniformly bounded. Now consider for instance p_n^+, and let $I_n^+ = [a,b]$. Then,

$$\text{(1.13)} \qquad \begin{aligned} ||p_n||_1 &\le \frac{1}{b-a}\left[\int_a^{\frac{a+b}{2}} \frac{x-a}{x}\, dx + \int_{\frac{a+b}{2}}^b \frac{b-x}{x}\, dx\right] \\ &\le \frac{1}{b-a}\left[a \ln\frac{2a}{a+b} + b \ln\frac{2b}{a+b}\right], \end{aligned}$$

which yields the desired result. The lemma is then proved.

Then $0 \le \hat{p}(\xi) < \infty$. Let ϕ be such that

$$\text{(1.14)} \qquad \hat{p}(\xi) = \left|\hat{\phi}(\xi)\right|^2 = \int_{|\xi|}^\infty \left|\hat{\psi}(u\,\mathrm{sgn}(\xi))\right|^2 \frac{du}{u}.$$

In other words, $|\hat{\psi}(u\xi)|^2 = -u\,\partial_u|\hat{\phi}(u\xi)|^2$ for all $\xi \in I\!R$, and $\lim_{\xi\to\infty}|\hat{\phi}(\xi)|^2 = 0$. ϕ is called a *scaling function*, and we associate to it the corresponding

$$\text{(1.15)} \qquad \phi_{(b,a)}(x) = \frac{1}{a}\,\phi\left(\frac{x-b}{a}\right).$$

We will see in §3 that $\phi(x)$ satisfies some kind of scaling equation (see (3.7)), and that infinite product formulas for $\hat{\phi}(\xi)$ can be obtained (3.5). Clearly, (1.14) does not characterize ϕ; although it is in general unnecessary, one can always restrict to a real-valued $\hat{\phi}$. Notice that the squared modulus of the Fourier transform of the scaling function is by construction a decreasing (respectively, increasing) function for positive (respectively, negative) values of ξ.

To any $f \in L^2(I\!R)$ associate

$$\text{(1.16)} \qquad s_a(x) = \int_{I\!R} \langle f, \phi_{(b,a)}\rangle \phi_{(b,a)}(x)\, db,$$

that is,

$$s_a(x) = \int_a^\infty d_u(x)\,\frac{du}{u} = (f * p_a)(x).$$

Then $s_a \in L^2(I\!R)$, and we have the following decompositions, whose proofs are immediate from that of Theorem 1.

COROLLARY. *Let ψ be an infinitesimal wavelet, and ϕ an associated scaling function. Then any $f \in L^2(I\!R)$ can be expressed as*

$$\text{(1.17)} \qquad f = \lim_{a\to 0} s_a$$

$$= s_{a_0} + \int_0^{a_0} d_a \frac{da}{a} \tag{1.18}$$

strongly in $L^2(\mathbb{R})$.

The corollary also holds in the $H^2(\mathbb{R})$ context.

In terms of linear filtering, ϕ can be seen as a low-pass filter, and ψ as a band-pass filter. Indeed, it is usual to consider analyzing wavelets such that $\hat{\psi}(\xi)$ is well localized in the Fourier domain around some frequency ω_0. Then $\hat{\psi}(t\xi)$ is localized around $\xi \cong \omega_0/t$, and $\hat{p}(\xi)$ (and $\hat{\phi}(\xi)$), built by "gluing together the $|\hat{\psi}(t\xi)|^2, t \in [1, \infty[$" is centered on the zero frequency. Then s_a describes the low-frequency content of f up to the scale a (otherwise stated as the approximation at scale a), and d_a describes the content of f around the scale a (i.e., the details at scale a).

Denote now by Q_a and P_a the convolution operators, defined, respectively, by the multipliers $\hat{q}_a(\xi) = |\hat{\psi}(a\xi)|^2$ and $\hat{p}_a(\xi) = |\hat{\phi}(a\xi)|^2$:

$$\left[\widehat{Q_a f}\right](\xi) = \hat{q}_a(\xi)\hat{f}(\xi) = \hat{d}_a(\xi), \tag{1.19}$$

$$\left[\widehat{P_a f}\right](\xi) = \hat{p}_a(\xi)\hat{f}(\xi) = \hat{s}_a(\xi). \tag{1.20}$$

The previous corollary then yields an approximation of the identity by the operators Q_a and P_a. If we introduce the spaces

$$V_{(a)} = P_a.L^2(\mathbb{R}),$$

then it clearly follows from the monotonicity of $|\hat{\phi}|$ that for any $a < a'$, $V_{(a')} \subseteq V_{(a)}$. By analogy with the usual multiresolution analysis [Me], we call such a collection of spaces a continuous (or infinitesimal) multiresolution analysis. In particular, it must be remarked that $V_{(a)}$ increases to the whole $L^2(\mathbb{R})$ as $a \to 0$ and decreases to $\{0\}$ as $a \to \infty$. Notice that all the $V_{(a)}$ spaces are translation invariant; then they are not closed, except in the case where $|\hat{\phi}| = \chi_\Omega \rho$, where χ_Ω is the characteristic function of some measurable set Ω and ρ a bounded strictly positive function (see, e.g., [Ru]).

1.2. Linear analysis. It is well known that the reconstructing and the analyzing wavelets can be decoupled. Otherwise stated one can use different infinitesimal wavelets for the computation of the coefficients and for the synthesis of the analyzed function from the coefficients (see [Ho-Tc], [Ho] for beautiful applications of this property in different contexts). In such a case, the admissibility condition (1.1) has to be modified accordingly [Ho-Tc]. A particular example of such a decoupling, which has been known for a long time, consists in taking formally a Dirac distribution for the reconstructing wavelet. Assuming instead of (1.1) that

$$k_\psi = \int_0^\infty \hat{\psi}(u) \frac{du}{u} = \int_0^\infty \hat{\psi}(-u) \frac{du}{u} = 1, \tag{1.21}$$

we have the following decomposition of any $f \in L^2(\mathbb{R})$:

$$f(x) = \int_{\mathbb{R}_+^*} \langle f, \psi_{(x,a)} \rangle \frac{da}{a} \tag{1.22}$$

strongly in $L^2(\mathbb{R})$. This is the so-called Morlet reconstruction formula of f from its wavelet coefficients. Such a linear analysis (linear in the ψ function) generates a continuous multiresolution analysis as follows: introduce the *linear scaling function* $\varphi \in L^1(\mathbb{R})$, defined by

$$\hat{\varphi}(\xi) = \int_{|\xi|}^{\infty} \hat{\psi}(u\,\mathrm{sgn}(\xi))\,\frac{du}{u}. \tag{1.23}$$

φ is also such that $\hat{\psi}(u\xi) = -u\,\partial_u\hat{\varphi}(u\xi)$ for all $\xi \in \mathbb{R}$. Associate to φ the following functions:

$$\varphi_{(b,a)}(x) = \frac{1}{a}\,\varphi\left(\frac{x-b}{a}\right). \tag{1.24}$$

Finally, introduce

$$\delta_a(x) = T_f(x,a) = \langle f, \psi_{(x,a)}\rangle \tag{1.25}$$

and

$$\sigma_a(x) = \langle f, \varphi_{(x,a)}\rangle. \tag{1.26}$$

We then have the linear analogue of Theorem 1 and the corresponding corollary.

THEOREM 1′. *Let $\psi \in H^1_a(\mathbb{R})$ be a mother wavelet, such that* (1.21) *holds, and let φ be the associated linear scaling function. Then any $f \in L^2(\mathbb{R})$ can be decomposed as*

$$\begin{aligned} f &= \lim_{a\to 0} \sigma_a && (1.27)\\ &= \sigma_{a_0} + \int_0^{a_0} \delta_a\,\frac{da}{a}, \qquad a_0 \in \mathbb{R}^*_+ && (1.28)\\ &= \int_0^{\infty} \delta_a\,\frac{da}{a} && (1.29)\end{aligned}$$

strongly in $L^2(\mathbb{R})$.

1.3. Comments and examples. The bilinear analysis gives the convenient scheme for the construction of orthonormal bases of wavelets [Me], [Da2]. Moreover, it is better adapted for the characterization of functional spaces (see, e.g., [Fr-Ja-We], [Ho-Tc]). On the other hand, the linear analysis is very often used for signal analysis [Gr-KM-Mo], since it produces a simplified reconstruction formula, involving a one-dimensional integral.

There are many examples of admissible infinitesimal wavelets, some of which are described in [Gr-Mo] and [Gr-KM-Mo]. Actually, the simplest one arises quite naturally from infinitesimal multiresolution analysis. Indeed, let us take a Gaussian scaling function:

$$\phi(x) = \frac{1}{\sqrt{\pi}}\,e^{-x^2},$$

where the normalization constant is fixed so that (1.1) holds. Then in the bilinear analysis scheme, the infinitesimal wavelets can be deduced from (1.7), and we can choose

$$\psi(x) = \frac{2x}{\sqrt{\pi}}\,e^{-x^2},$$

i.e., a derivative of Gaussian. The case of linear analysis yields a more famous infinitesimal wavelet. Taking for φ a Gaussian function leads to the celebrated Laplacian of Gaussian, used for a long time in image analysis and processing, and vision (see, e.g., [DD], [Mar]):

$$\psi(x) = \frac{1}{\sqrt{\pi}} \, (1 - 2x^2) \, e^{-x^2}.$$

In the next section, we will see how such derivatives of Gaussians simply lead to differences of Gaussians, which are also used in image theory and vision.

2. Wavelet packets in $L^2(I\!R)$. An important problem is that of the discretization of the continuous formulas derived in §1. At least for numerical applications, we want to be able to get discrete approximations of the identity in $L^2(I\!R)$, and to control the discretization error. Such a problem was handled in [Dal], where the author developed the theory of frames of wavelets. In particular, focusing on the scale discretization problem: if $\Lambda = \{a_0\lambda_0^n, n \in Z\!\!Z\}$ denotes a geometric sequence in $I\!R$, for some positive λ_0, then $\sum_{\lambda\in\Lambda} Q_\lambda$ defines a linear operator, which is positive, bounded, and invertible with bounded inverse for a suitable choice of the λ_0 parameter and the infinitesimal wavelet ψ. Moreover, this operator can often be written as a small perturbation of the identity (up to some multiplicative constant), which permits us to consider the discretization as a good approximation of the continuous formula.

We are interested here in another way of discretizing the scales, such that the discretization errors are avoided. Such a procedure canonically produces new waveforms that we will call wavelet packets. Start from a strictly decreasing sequence of positive real numbers:

$$\cdots < a_{j+1} < a_j < a_{j-1} < \cdots$$

such that $\lim_{j\to-\infty} a_j = \infty$ and $\lim_{j\to+\infty} a_j = 0$.

Set

$$D_j(x) = \int_{a_{j+1}}^{a_j} d_a(x) \, \frac{da}{a} \, . \tag{2.1}$$

Then $D_j \in L^2(I\!R)$, and

$$\widehat{D_j}(\xi) = \hat{f}(\xi) \int_{a_{j+1}}^{a_j} \left|\hat{\psi}(a\xi)\right|^2 \frac{da}{a} \quad \text{almost everywhere.} \tag{2.2}$$

Introducing the function Ψ^j, such that

$$\left|\widehat{\Psi^j}(\xi)\right|^2 = \int_{a_{j+1}}^{a_j} \left|\hat{\psi}(a\xi)\right|^2 \frac{da}{a}, \tag{2.3}$$

we then have

$$\widehat{D_j}(\xi) = \hat{f}(\xi) \, \left|\widehat{\Psi^j}(\xi)\right|^2 . \tag{2.4}$$

We will refer to the Ψ^j functions as *wavelet packets*, since they are built up by gluing wavelets together (in the Fourier space, contrary to the case of those described in [Co-M-Q-W]). Such wavelet packets are also intuitively close to the atoms introduced in Littlewood–Paley theory (see, e.g., [Fr-Ja-We]), obtained via a segmentation of Calderón's formula (1.3) into integrals over dyadic cubes. In our case, the wavelet packets are obtained by a segmentation of (1.3) into integrals over strips in the time-scale plane. The result then is that the atoms are generated in a simple way. Notice that (2.3) does not completely define the wavelet packets. Once again, we can restrict to wavelet packets with positive-valued Fourier transform, but this is not necessary. By construction, the wavelet packets lead to a partition of unity in the Fourier space as follows:

$$\sum_{j=-\infty}^{+\infty} \left|\widehat{\Psi^j}(\xi)\right|^2 = \left|\widehat{\Phi^{j_0}}(\xi)\right|^2 + \sum_{j=j_0}^{\infty} \left|\widehat{\Psi^j}(\xi)\right|^2 = 1 \tag{2.5}$$

for all $\xi \in I\!R$, where we have set

$$\widehat{\Phi^j}(\xi) = \hat{\phi}(a_j\xi). \tag{2.6}$$

Defining the translates of the Φ^j and Ψ^j as $\Phi^j_b(x) = \Phi^j(x-b)$ and $\Psi^j_b(x) = \Psi^j(x-b)$, we then have the following.

THEOREM 2. *Let ψ be an infinitesimal wavelet, and let Ψ^j and Φ^j be associated wavelet packets and scaling functions as in* (2.3) *and* (2.6). *Then any $f \in L^2(I\!R)$ can be decomposed as*

$$f = \int_{I\!R} \langle f, \Phi^{j_0}_b\rangle\, \Phi^{j_0}_b\, db + \sum_{j=j_0}^{\infty} \int_{I\!R} \langle f, \Psi^j_b\rangle\, \Psi^j_b\, db \tag{2.7}$$

strongly in $L^2(I\!R)$.

Denote by $\mathbf{Q}_j$ the convolution operator, defined by the multiplier $|\widehat{\Psi^j}(\xi)|^2$, and set $\mathbf{P}_j = P_{a_j}$. Let V_j and W_j denote, respectively, the images of $L^2(I\!R)$ by $\mathbf{P}_j$ and $\mathbf{Q}_j$. This provides the following resolution of $L^2(I\!R)$:

$$\cdots \subset V_{j-1} \subset V_j \subset V_{j+1} \subset \cdots. \tag{2.8}$$

Again, it must be remarked that V_j increases to the whole $L^2(I\!R)$ as $j \to \infty$ and decreases to $\{0\}$ as $j \to -\infty$. Moreover, $V_j + W_j = V_{j+1}$, but in general $V_i \not\perp W_i$.

Remarks. The wavelet packets have vanishing integral by construction. Moreover, they can be seen as differences of two low-pass filters. Indeed, we have

$$\left|\widehat{\Psi^j}(\xi)\right|^2 = \left|\widehat{\Phi^{j+1}}(\xi)\right|^2 - \left|\widehat{\Phi^j}(\xi)\right|^2 \tag{2.9}$$

and the partial reconstructions

$$D^j(x) = \int_{I\!R} \langle f, \Psi^j_b\rangle\, \Psi^j_b(x)\, db \tag{2.10}$$

as differences of two smoothings of $f(x)$ at two consecutive scales:

$$D^j(x) = S^{j+1}(x) - S^j(x), \tag{2.11}$$

where

$$S^j(x) = \int_{I\!R} \langle f, \Phi^j_b\rangle\, \Phi^j_b(x)\, db. \tag{2.12}$$

Once more, we are close to the idea of the "difference of two smoothings" wavelet familiar in vision theory.

The same can be done in the linear analysis scheme. The wavelet packets are then defined as

$$\widehat{\Theta^j}(\xi) = \int_{a_{j+1}}^{a_j} \hat{\psi}(a\xi)\,\frac{da}{a} \tag{2.13}$$

and yield a partition of unity in the Fourier space:

$$\sum_{j=-\infty}^{+\infty} \widehat{\Theta^j}(\xi) = \hat{\varphi}(a_{j_0}\xi) + \sum_{j=j_0}^{\infty} \widehat{\Theta^j}(\xi) = 1. \tag{2.14}$$

The linear wavelet packets still appear as differences of smoothings at two consecutive scales, as

$$\widehat{\Theta^j}(\xi) = \hat{\varphi}(a_{j+1}\xi) - \hat{\varphi}(a_j\xi) \tag{2.15}$$

and every $f \in L^2(I\!R)$ decomposes as

$$f(x) = \sum_{j=-\infty}^{\infty} \langle f, \Theta^j_x \rangle, \tag{2.16}$$

where $\Theta^j_b(x) = \Theta^j(x-b)$.

The Gaussian example is once more very interesting, since it provides directly the DOGs, i.e., the differences of Gaussians at two different scales (see, e.g., [Mar]).

Consider the particular case where all the scale parameters a_j are generated from a unique one a_0 as a geometric sequence:

$$a_j = a_0\lambda^{-j}$$

for some positive real number $\lambda > 1$. Then the wavelet packets Ψ^j (and also the Θ^j) can be considered as wavelets in the usual sense, since they are generated from a unique function Ψ^0 (or Θ^0) by dilations by powers of λ. In particular, the bilinear analysis yields the usual Littlewood–Paley analysis (also called dyadic wavelet analysis in [Ma]).

Finally, let us stress that although we have chosen to use the term "Wavelet Packets," our construction is completely different than the wavelet packets construction proposed in [Co-M-Q-W]. Both constructions share the fact that the corresponding functions are of variable bandwidth, and that the bandwidth can be matched to a function to be analyzed.

3. Mirror filters and factorizable wavelet packets.

3.1. Mirror filters. The orthonormal bases of wavelet packets described, for example, in [Co-M-Q-W] are derived in a very simple algorithmic way from the so-called Quadrature Mirror Filters (QMFs), canonically associated with the multiresolution analysis. We will see in this section how infinitesimal multiresolution analysis can produce continuous versions of such filters, and how these infinitesimal filters can be put together to form QMFs for the wavelet packets. Let us start from the bilinear infinitesimal multiresolution analysis (the same arguments can be developed in the linear analysis scheme, too), with scaling function ϕ. If $|\hat{\phi}(a\xi)| \neq 0$, it follows from the monotonicity property of $|\hat{\phi}|$ that a determination of the complex logarithm $\ln\hat{\phi}(u\xi)$ can always be chosen for $0 \leq u \leq a$, so that the function

$$\nu_0(a\xi) = a\partial_a \ln\hat{\phi}(a\xi) \quad \forall a, \xi \quad \text{such that } \hat{\phi}(a\xi) \neq 0 \tag{3.1}$$

can be introduced. Then clearly

$$\ln\hat{\phi}(a\xi) = \int_0^a \nu_0(u\xi)\,\frac{du}{u} \tag{3.2}$$

and

$$\hat{\phi}(\xi) = \exp\left[\int_0^1 \nu_0(u\xi)\,\frac{du}{u}\right]. \tag{3.3}$$

Such a formula can be thought of as a continuous multiplicative formula, interpreting the exponential of the integral as a continuous product of exponentials. Indeed, setting

$$\mu_0(\xi) = e^{\nu_0(\xi)}, \tag{3.4}$$

we can write the Fourier transform of the scaling function as

$$\hat{\phi}(\xi) = \prod_{u=0}^{1}\left(\frac{du}{u}\right)\,\mu_0(u\xi), \tag{3.5}$$

the continuous product $\prod_{u=0}^{1}\left(\frac{du}{u}\right)$ with respect to the logarithmic measure $\frac{du}{u}$ being defined by the right-hand side of (3.3).

For ξ such that $\hat{\phi}(a\xi) \neq 0$, introduce now the functions $m_0^j(\xi)$, called the *integrated filters*:

$$m_0^j(\xi) = \prod_{u=a_{j+1}}^{a_j}\left(\frac{du}{u}\right)\,\mu_0(u\xi). \tag{3.6}$$

Such functions allow a discretization of the continuous product formula (3.5), and make the connection with the structure appearing in the usual multiresolution context. The Fourier transform of the scaling functions can be written as

$$\hat{\phi}(a_j\xi) = m_0^j(\xi)\hat{\phi}(a_{j+1}\xi). \tag{3.7}$$

Notice that this expression, together with the monotonicity of $|\hat{\phi}|$, can be used to define $m_0^j(\xi)$, according to

$$m_0^j(\xi) = \begin{cases} \frac{\hat{\phi}(a_j\xi)}{\hat{\phi}(a_{j+1}\xi)} \quad \forall\,\xi \quad \text{such that } \hat{\phi}(a_{j+1}\xi) \neq 0, \\ 0 \quad \forall\,\xi \quad \text{such that } \hat{\phi}(a_{j+1}\xi) = 0. \end{cases} \tag{3.8}$$

Introducing now the high-pass filter

$$m_1^j(\xi) = \begin{cases} \frac{\hat{\Psi}^j(\xi)}{\hat{\phi}(a_{j+1}\xi)} \quad \forall\,\xi \quad \text{such that } \hat{\phi}(a_{j+1}\xi) \neq 0, \\ 0 \quad \forall\,\xi \quad \text{such that } \hat{\phi}(a_{j+1}\xi) = 0, \end{cases} \tag{3.9}$$

an immediate consequence of (2.9) is that the filters m_0^j and m_1^j fulfill the quadrature mirror filter condition, that is,

$$\left|m_0^j(\xi)\right|^2 + \left|m_1^j(\xi)\right|^2 = 1 \ \ \forall j \in \mathbb{Z}, \ \ \forall \xi \in \mathbb{R} \ \ \text{such that } \hat{\phi}(a_{j+1}\xi) \neq 0. \tag{3.10}$$

3.2. Factorizable wavelet packets. We will call *factorizable wavelet packets* the wavelet packets generated by an infinitesimal wavelet ψ and a decreasing sequence $\{a_j\}$ of positive real numbers with the property that there exists a positive real $\lambda > 1$ such that for all $j \in \mathbb{Z}$,

$$a_{j+1} = \lambda^{-n(j)} a_j \tag{3.11}$$

for some $n(j) \in \mathbb{N}^*$. In other words, factorizable wavelet packets correspond to sequences of scale parameters that are subsequences of geometric sequences. Introduce then the filter m_0 such that

$$\hat{\phi}(\lambda\xi) = m_0(\xi)\hat{\phi}(\xi), \tag{3.12}$$

that is,

$$m_0(\xi) = \prod_{u=1}^{\lambda} \left(\frac{du}{u}\right) \ \mu_0(u\xi) = e^{\int_1^\lambda \nu_0(u\xi) \, \frac{du}{u}}. \tag{3.13}$$

In that case, all the integrated filters m_0^j factorize into products of dilated copies of m_0:

$$m_0^j(\xi) = \prod_{n=0}^{n(j)-1} m_0(\lambda^n a_j \xi). \tag{3.14}$$

This means that for numerical computations, the same filter can be used throughout the decomposition. It is then possible to build an algorithm for the computation of the wavelet packets coefficients with a pyramidal structure, as in the case of the QMF algorithm for the computation of the coefficients with respect to an orthonormal base of wavelets. Factorizable wavelet packets then seem more adapted for numerical implementation. This does not mean of course that nonfactorizable schemes are impossible to implement. There can be some specific problems (namely, in acoustics and voice analysis, or in vision) for which the optimal paving of the Fourier space is fixed by some phenomenological results, and do not a priori correspond to factorizable schemes. Nevertheless, depending on the required precision, it is likely that such pavings can be approached by pavings corresponding to factorizable schemes.

Remark. In the particular case where $n(j)$ is constant (and can be set to 1 without loss of generality), all the wavelet packets are obtained by dilations of a unique function Ψ, and all the integrated filters are dilated copies of m_0, too. The previous discussion then leads to a fast (i.e., $N \ln N$) algorithm, with pyramidal structure, to compute continuous wavelet decompositions. Such an algorithm is in the same spirit as the "Algorithme à Trous" discussed in [Ho-KM-T], since it is obtained by replacing the initial infinitesimal wavelet by another wavelet for which we can use QMFs. However, while the "Algorithme à Trous" is associated to an interpolation scheme, and then damages the localization of the wavelets in the Fourier space (see, e.g., [Du]), the one proposed here uses the function Ψ, which has essentially the same decay properties as ψ in the Fourier space. It then seems more adapted to signal analysis, at least for applications like those described in [Gr-KM-Mo]. Let us nevertheless stress that we have not proposed any discretization scheme for the (continuous) variations of the wavelet packets coefficient with respect to the translation parameter b. A brutal discretization of the translation parameter periodises the m_0 and then the m_1 filters, yielding a more usual QMF relation. However, the exact reconstruction property is then lost (notice that the errors can in general be controlled).

4. Factorization formulas and contrast analysis. We now describe a multiplicative way of reconstructing a function $f \in L^2(I\!R)$ from s_a and the contrast functions, which can be introduced as $c_a = \frac{d_a}{s_a}$. A precise meaning can be given to the multiplicative reconstruction formula, by making some analyticity assumptions on the scaling function ϕ, ensuring that the zeros of $s_a(x)$ are isolated (with respect to the dilation variable a, and for any $x \in I\!R$). We will discuss this analyticity property at the end of this section.

Let us first assume that s_a is a real analytic function of $a \in]0, \infty[$. Then, if $s_u(x)$ does not vanish for $u \in [a, A]$, we can write

$$s_a(x) = s_A(x)\; e^{-\int_a^A \partial_u \ln s_u(x)\, du}.$$

Now if $\partial_u s_u(x) = -d_u(x)/u$ (which is right under the assumptions made on the scaling function ϕ), we then have

$$s_a(x) = s_A(x)\; e^{\int_a^A \frac{d_u(x)}{s_u(x)} \frac{du}{u}}. \tag{4.1}$$

Let us now make sense of the integral when $s_u(x)$ vanishes in $[a, A]$. The goal is that (4.1) still holds in such a case. Let us assume that a_0 is the unique zero of $s_u(x)$ in the interval $[a, A]$, and set, for $\varepsilon > 0$,

$$\begin{aligned} s_a(x) &= \frac{s_a(x)}{s_{a_0-\varepsilon}(x)} \frac{s_{a_0-\varepsilon}(x)}{s_{a_0+\varepsilon}(x)} \frac{s_{a_0+\varepsilon}(x)}{s_A(x)} s_A(x) \\ &= s_A(x)\; \exp\left[\left(\int_a^{a_0-\varepsilon} \frac{du}{u} + \int_{a_0+\varepsilon}^A \frac{du}{u}\right) \frac{d_u(x)}{s_u(x)}\right] \frac{s_{a_0-\varepsilon}(x)}{s_{a_0+\varepsilon}(x)}. \end{aligned}$$

Taking the limit $\varepsilon \longrightarrow 0$, the analyticity of $s_u(x)$ implies the existence of $k \in I\!N$ and a function h such that $s_u(x) = (u - a_0)^k h(u, x)$, with $h(a_0, x) \neq 0$. Then

$$\lim_{\varepsilon \to 0} \frac{s_{a_0-\varepsilon}(x)}{s_{a_0+\varepsilon}(x)} = (-1)^k \quad \text{and} \quad \frac{d_u(x)}{s_u(x)} = \frac{k}{u - a_0} + g(u, x),$$

for some function g continuous at $u = a_0$, otherwise stated as

$$s_a(x) = s_A(x)\; \exp\left[\text{principal value} \int_a^A \frac{d_u(x)}{s_u(x)} \frac{du}{u} \;+\; k\pi i\right]. \tag{4.2}$$

Notice that such a specification of the integral coincides with the one obtained by redefining the (real) integration interval $a \leq u \leq A$ by the union of the two intervals $a \leq u \leq a_0 - \varepsilon$, $a_0 - \varepsilon \leq u \leq A$, and the half-circle of radius ε centered at a_0, and then taking the limit $\varepsilon \to 0$. Notice also that when $a_0 = a$ or $a_0 = A$, the integral diverges, which is compatible with (4.1).

Let us denote by Ξ the set of admissible infinitesimal wavelets such that in addition for any $f \in L^2(\mathbb{R})$, $s_a(x)$ is analytic with respect to a for any $x \in \mathbb{R}$. We can then introduce the *infinitesimal contrast function*

$$c_a(x) = -a\partial_a \ln s_a(x) \ \ \forall a, x \quad \text{such that } s_a(x) \neq 0 \tag{4.3}$$

$$= \frac{d_a(x)}{s_a(x)}, \tag{4.4}$$

that is, essentially the details of $f(x)$ at the scale a divided out by the approximation of $f(x)$ at the scale a (this explains the name of contrast function). Then we have shown the following.

THEOREM 3. *Let $\psi \in \Xi$ be an infinitesimal wavelet, and let $f \in L^2(\mathbb{R})$. Then, for any $a < A$, we have*

$$s_a(x) = s_A(x) \ \exp\left[\int_a^A c_u(x)\,\frac{du}{u}\right] \tag{4.5}$$

with the above specification of the integral.

Now (4.4) can also be written as a multiplicative integral, or continuous product as follows:

$$s_a(x) = s_A(x) \prod_{u=a}^{A} \left(\frac{du}{u}\right) e^{c_u(x)}, \tag{4.6}$$

the continuous product being defined by (2.5).

Remark. The factorization formula is independent of the determination of the complex logarithm between two singularities. It is then possible to specify a global determination of the logarithm for $a < u < A$, turning around the singularities.

Let us now come back to the question of the analyticity of $s_a(x)$. We have the following.

LEMMA. *If $|\hat{\phi}(\xi)|^2$ is infinitely differentiable, such that its derivatives satisfy the bound*

$$|\partial_\xi^k |\hat{\phi}(\xi)|^2| \ \leq \ K C^k k!\,(1+|\xi|)^{-1-k} \qquad \forall \xi \in \mathbb{R},\ k \in \mathbb{N} \tag{4.7}$$

for some constants $C, K > 0$, then for any $f \in L^2(\mathbb{R})$, $s_a(x)$ is an analytic function of $a \in \mathbb{R}_+^*$ for all $x \in \mathbb{R}$. Moreover, $\partial_a s_a(x) = -\frac{1}{a} d_a(x)$.

Remark. Actually, such assumptions imply the strong (i.e., in norm) analyticity of the map $a \to |\hat{\phi}(a\xi)|^2$ with values in $L^2(\mathbb{R})$.

Proof of the lemma. Set $h(\xi) = |\hat{\phi}(\xi)|^2$. Let us estimate the remainder of the Taylor series of $s_a(x)$ around $a = a_0$:

$$r_n = \frac{(a-a_0)^n}{n!}\ \partial_a \left[\int \hat{f}(\xi)\, h(a\xi)\xi^n d\xi\right]_{a=\tilde{a}_0} \tag{4.8}$$

for $\tilde{a}_0$ between a and a_0. By assumption, we can find some μ such that

$$\int \sup_{|a-\tilde{a}_0|<\mu} \left|\hat{f}(\xi)\,(\partial^n h)(a\xi)\xi^n\right| d\xi < \infty,$$

so that the integral and the derivative can be permuted. Then

$$|r_n| \leq |a-a_0|^n \; K C^n \int \left|\hat{f}(\xi)\right| \left(1+|\tilde{a}_0\xi|\right)^{-1-n} d\xi. \tag{4.9}$$

Then $\lim_{n\to\infty} \sup_{|a-\tilde{a}_0|<\epsilon} |r_n| = 0$ for $|a-a_0| < \epsilon$ and ϵ small enough, which proves the lemma.

It is worth noticing that contrast analysis becomes particularly simple in the case of the analysis of positive-valued functions, and positive-valued scaling functions (such as Gaussian functions for instance), as observed in [DD]. Indeed, in such cases one does not have to take care of the zeros of the s_a functions.

Let us finally briefly describe the discretization of the factorization formula, more precisely its relationship to the wavelet packets we described in §§2 and 3. Consider again the strictly decreasing sequence a_j of positive real numbers. Then for all integers $n > m$ the integral $\int_{a_m}^{a_n} c_u(x)\,\frac{du}{u}$ can be truncated into integrals over smaller intervals, and the factorization becomes

$$s_{a_n}(x) = s_{a_m}(x) \prod_{j=m}^{n-1} \exp\left[\int_{a_{j+1}}^{a_j} c_u(x)\,\frac{du}{u}\right],$$

which reduces to the trivial expression

$$s_{a_n}(x) = s_{a_m}(x) \prod_{j=m}^{n-1} \frac{s_{a_{j+1}}(x)}{s_{a_j}(x)}. \tag{4.10}$$

Such a simple factorization can still be expressed in terms of *integrated contrast coefficients*. Indeed, introducing a family of wavelet packets such that (2.3) holds, the associated integrated contrast coefficients can be defined as

$$C_j(x) = \frac{D_j(x)}{s_{a_j}(x)}, \tag{4.11}$$

so that the factorization formula reads

$$s_{a_n}(x) = s_{a_m}(x) \prod_{j=m}^{n-1} \left(1 + C_j(x)\right) . \tag{4.12}$$

The contrast coefficients form a sufficient information for the characterization of the analyzed function. The scheme defined by (4.8) can very easily be used for numerical computations [DD], [Me-DD-Ge].

Remark. Linear contrast analysis: Let us finally describe the linear analogue of the previous contrast analysis. Owing to Theorem 1′, we introduce the contrast function

$$\chi_a(x) = \frac{\delta_a(x)}{\sigma_a(x)}. \tag{4.13}$$

Then, assuming that $\hat{\varphi}$ is of class C^∞, such that in addition

$$\left|\partial_\xi^k |\hat{\varphi}(\xi)|\right| \leq K C^k k! (1+|\xi|)^{-1-k} \quad \forall \xi \in I\!R,\ k \in I\!N \tag{4.14}$$

for some positive constants C, K, we obtain the analyticity of $\sigma_a(x)$ and

$$\sigma_a(x) = \sigma_A(x) \prod_{u=a}^{A} \left(\frac{du}{u}\right) e^{\chi_u(x)}. \tag{4.15}$$

Such a continuous formula can also be discretized along the same lines as (4.6).

5. Conclusions. The continuous wavelet transform is naturally associated with an infinitesimal multiresolution scheme. The wavelet transform and the continuous set of approximations of the analyzed function are linked by some kind of scale derivative. This relationship is used to build a discrete set of wavelet packets by partial integration of the infinitesimal wavelet on scale intervals. Such wavelet packets allow exact reconstruction, though the scale axis has been discretized. If the sequence of scale parameters involved is factorizable as a subsequence of a geometric one, the scale discretization scheme can be combined with a pyramidal time discretization scheme. We then lose the exact reconstruction property (notice that the corresponding discretization errors can probably be controled through appropriate frame estimates). We have shown that this way of discretizing an infinitesimal multiresolution scheme on the scales uses the wavelet transform as a scale derivative. Moreover, it may be generalized to a logarithmic derivative. The resulting infinitesimal analysis is the "contrast function," and leads to a multiplicative reconstruction formula. Such a logarithmic way of performing multiscale analysis is, in particular, of great interest for the analysis of positive-valued signals, for which the troubles due to the zero-crossings are avoided. It may be extended directly to the two-dimensional case, and applied to image analysis. It then provides a rigorous framework to associate two basic properties of human vision: multiscale information processing and logarithmic light sensitivity [Me-DD-Ge]. The discrete reconstruction scheme, based on partial integrations of contrast functions, gives rise to formulas that are almost equivalent to "ratios of low pass filters" (ROLPs) decompositions that have already proved their efficiency in image processing [To-Ru-Va].

Acknowledgments. We are very indebted to Ph. Tchamitchian for many enlightening discussions, and particularly for showing us that $p \in L^1(I\!R)$. We thank D. J. Ripmeester for a careful reading of the manuscript. Thanks are also due to A. Grossmann and Y. Meyer. M. A. Muschietti is grateful to the Centre National de la Recherche Scientifique, the Universities of Aix-Marseille I and III, and the University of Nancy for kind invitations.

REFERENCES

[Cal] A. Calderón, *Intermediate spaces and interpolation, the complex method*, Studia Math., 24 (1964), p. 113.

[Co-M-Q-W] R. Coifman, Y. Meyer, S. Quake, and M. V. Wickerhauser, *Signal processing and compression with wavelet packets*, Yale University, New Haven, CT, preprint, 1990.

[Co-We] R. Coifman and G. Weiss, *Extensions of Hardy spaces and their use in analysis*, Bull. Amer. Math. Soc., 83 (1977), p. 569.

[Co-Gr-Tc] J. M. Combes, A. Grossmann, and Ph. Tchamitchian, eds., *Wavelets, time-frequency analysis and phase space*, Proceedings of an international conference, Marseille, Inverse Probl. Theoret. Imaging, Springer-Verlag, Berlin, 1989.

[Da1] I. Daubechies, *The wavelet transform, time-frequency analysis, and signal analysis*, IEEE Trans. Inform. Theory, 36 (1990), p. 961.

[Da2] ———, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[De] N. Delprat, B. Escudié, P. Guillemain, R. Kronland-Martinet, Ph. Tchamitchian, and B. Torrésani, *Asymptotic wavelet and Gabor analysis: extraction of instantaneous frequencies*, IEEE Trans. Inform. Theory, 38 (1992), pp. 644–664.

[Du] P. Dutilleux *An implementation of the "Algorithme à Trous" to compute the wavelet transform*, in Wavelets, Time-Frequency Methods and Phase Space, Combes, et al., eds., Inverse Probl. Theoret. Imaging, Springer-Verlag, Berlin, 1989, p. 298.

[DD] M. Duval-Destin, *Analyse spatio-temporelle de la simulation visuelle à l'aide de la transformée en ondelettes*, Ph.D. thesis, Université d'Aix–Marseille II, 1991.

[Fr-Ja-We] M. Frazier, B. Jawerth, and G. Weiss, *Littlewood–Paley Theory and the Study of Function Spaces*, CBMS-NSF Regional Conf. Ser. in Appl. Math., to appear.

[Fr-Mo] J. Froment and J. M. Morel, *Analyse multiéchelle, vision stéréo et ondelettes*, in Les ondelettes en 1989, P.G. Lemarié, ed., Lecture Notes in Math., Springer-Verlag, Berlin, 1990.

[Ga] D. Gabor, *Theory of communication*, J. I.E.E. (London), 93 (1946), p. 429.

[Gr-KM-Mo] A. Grossmann, R. Kronland-Martinet, J. Morlet, *Reading and understanding continuous wavelet transform*, in Wavelets, Time-Frequency Methods and Phase Space, Combes et al., eds., Inverse Probl. Theoret. Imaging, Springer-Verlag, Berlin, 1989, p. 2.

[Gr-Mo] A. Grossmann and J. Morlet, *Decomposition of Hardy functions into wavelets of constant shape*, SIAM J. Math. Anal., 15 (1984), p. 783.

[Gui] A. Guichardet, *Symmetric Hilbert spaces and related topics*, Lecture Notes in Math. 261, Springer-Verlag, Berlin, 1972.

[He-Wa] C. Heil and D. Walnut, *Continuous and discrete wavelet transform*, SIAM Rev., 31 (1989), pp. 628–666.

[Ho] M. Holschneider, *Inverse Radon transform through inverse wavelet transform*, preprint CPT-90/P2364; Inverse Problems, submitted.

[Ho-KM-T] M. Holschneider, R. Kronland-Martinet, J. Morlet, and Ph. Tchamitchian, *A real time algorithm for signal analysis with the help of the wavelet transform*, in Wavelets, Time-Frequency Methods and Phase Space, J. M. Combes et al., eds., Inverse Probl. Theoret. Imaging, Springer-Verlag, Berlin, 1989, p. 286.

[Ho-Tc] M. Holschneider and Ph. Tchamitchian, *Pointwise analysis of Riemann's non-differentiable function*, Invent. Math., 105 (1991), p. 157.

[Ma] S. Mallat, *Multiresolution approach to wavelets in computer vision*, in Wavelets, Time-Frequency Methods and Phase Space, J. M. Combes et al., eds., Inverse Probl. Theoret. Imaging, Springer-Verlag, Berlin, 1989, p. 298.

[Mar] D. Marr, *Vision*, Freeman, San Francisco, CA, 1982.

[Me-DD-Ge] J. P. Menu, M. Duval-Destin, and T. Gervais, *Spatial metric and image processing: an opportunity for advanced displays development*, Proc. of the Conf. of the Society for Information Display, Las Vegas, NV, May 1990, p. 468.

[Me] Y. Meyer, *Ondelettes et Opérateurs,* I *Ondelettes*, Hermann, Paris, 1989.

[Ru] W. Rudin, *Real and Complex Analysis*, Second Ed., McGraw-Hill, 1987.

[Str] J. O. Stromberg, *A modified Haar system and higher order spline systems*, Conference in Harmonic Analysis in Honor of Atoni Zygmund, II, W. Beckner et al., eds., Wadworth Math. Series, p. 475.

[To-Ru-Va] A. Toet, L. Ruyven, and J. Valeton, *Merging thermal images by a contrast pyramid*, Opt. Engrg., 28 (1989), p. 789.

[To1] B. Torrésani, *Wavelets associated with representations of the affine Weyl–Heisenberg group*, J. Math. Phys., 32 (1991), p. 1273.

[To2] ———, *Time-frequency representations: wavelet packets and optimal decompositions*, Ann. Inst. H. Poincaré, Phys. Théor., 56 (1992), pp. 215–234.

[Wi] M. V. Wickerhauser, INRIA *lectures on wavelet packets algorithms*, Yale University, New Haven, CT, preprint, 1991.

SIAM J. MATH. ANAL.
Vol. 24, No 3, pp. 756-767, May 1993

013

EXPONENTIALLY-IMPROVED ASYMPTOTIC SOLUTIONS OF ORDINARY DIFFERENTIAL EQUATIONS I: THE CONFLUENT HYPERGEOMETRIC FUNCTION*

F. W. J. OLVER†

Abstract. There has been renewed interest in both formal and rigorous theories of exponentially-small contributions to asymptotic expansions. In particular, a generalized asymptotic expansion was obtained for the confluent hypergeometric function $U(a, a-b+1, z)$ in which the parameters a and b are complex constants, and z is a large complex variable. This expansion is expressed in terms of generalized exponential integrals and has a larger region of validity and greater accuracy than conventional expansions of Poincaré type. The expansion was established by transformations and a re-expansion of an integral representation of $U(a, a-b+1, z)$. In this paper it is shown how the same result can be achieved by a direct differential-equation approach, thereby laying the foundation for a rigorous theory of generalized asymptotic solutions of linear differential equations.

Key words. factorial series, Gamma function, generalized asymptotic expansion, generalized exponential integral, incomplete Gamma function, superasymptotics

AMS subject classifications. primary 34E05; secondary 33C15

1. Introduction. The confluent hypergeometric function $U(a, a-b+1, z)$ has the well-known asymptotic expansion

$$U(a, a-b+1, z) \sim z^{-a} \sum_{s=0}^{\infty} (-)^s \frac{(a)_s (b)_s}{s!\, z^s}, \tag{1.1}$$

valid when the real or complex parameters a and b are fixed, and the argument $z \to \infty$ in the sector $|\mathrm{ph}\, z| \leqq \frac{3}{2}\pi - \delta$. Here $(a)_s$ and $(b)_s$ denote the ascending factorials $a(a+1)\cdots(a+s-1)$ and $b(b+1)\cdots(b+s-1)$, respectively, and δ is an arbitrary positive constant. If either a or b is a nonpositive integer, then the expansion (1.1) terminates and furnishes an exact representation of $U(a, a-b+1, z)$. In other cases, for given values of a, b, and z, $|z|$ being sufficiently large, the terms decrease numerically to a minimum; thereafter they increase unboundedly. The accuracy obtainable from (1.1) is limited, and greatest when the summation ceases close to the numerically smallest term. Approximately, this term is given by $s = \mathrm{int}\,[|z|]$, the integer part of $|z|$, or somewhat more exactly by $s = \mathrm{int}\,[|z| - \mathrm{Re}\, a - \mathrm{Re}\, b + 2]$.

In a recent paper [7] the author established the following result.

THEOREM 1.1. *Define $R_n(a, b, z)$ by*

$$U(a, a-b+1, z) = z^{-a} \sum_{s=0}^{n-1} (-)^s \frac{(a)_s (b)_s}{s!\, z^s} + R_n(a, b, z), \tag{1.2}$$

where

$$n = |z| - a - b + 1 + \alpha, \tag{1.3}$$

$|z|$ being large, a and b being fixed real or complex parameters, and $|\alpha|$ being bounded.

* Received by the editors April 15, 1992; accepted for publication (in revised form) August 19, 1992.

† Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742.

This research was supported by National Science Foundation grant DMS 87-23039 and by a Senior Visiting Fellowship from the Royal Irish Academy.

Then

$$R_n(a,b,z)=(-)^n 2\pi\frac{e^z z^{b-1}}{\Gamma(a)\Gamma(b)}\left\{\sum_{s=0}^{m-1}(-)^s\frac{(1-a)_s(1-b)_s}{s!}\frac{F_{n+a+b-s-1}(z)}{z^s}\right.$$
$$\left.+(1-a)_m(1-b)_m R_{m,n}(a,b,z)\right\}, \tag{1.4}$$

where m is an arbitrary fixed nonnegative integer,

$$R_{m,n}(a,b,z)=O(e^{-z-|z|}z^{-m}),\qquad |\text{ph } z|\leqq\pi, \tag{1.5}$$

$$R_{m,n}(a,b,z)=O(z^{-m}),\qquad \pi\leqq|\text{ph } z|\leqq\tfrac{5}{2}\pi-\delta, \tag{1.6}$$

and δ denotes an arbitrary positive constant.

In this result F denotes the function defined by

$$F_p(z)=\frac{e^{-z}}{2\pi}\int_0^\infty\frac{e^{-zt}t^{p-1}}{1+t}\,dt, \tag{1.7}$$

when Re $p>0$ and $|\text{ph } z|<\frac{1}{2}\pi$, and by analytic continuation elsewhere.

It will be observed that Theorem 1.1 provides a re-expansion of the remainder term $R_n(a, b, z)$ when the original expansion (1.1) is truncated at or near its optimal stage; compare (1.3). In applications there are two advantages of the expansion given by (1.2) and (1.4), compared with (1.1). First, the total sector of validity $|\text{ph } z|\leqq\frac{5}{2}\pi-\delta$ is greater. Secondly, the attainable accuracy is considerably improved.

Another important feature of Theorem 1.1 is that it leads to a very clear understanding of the Stokes phenomenon: indeed, it was the insightful formal researches of Berry [1] on this aspect that motivated the research leading to Theorem 1.1. In brief, the explanation is that if n and $|z|$ are large, fixed, and approximately equal, then the functions $F_{n+a+b-s-1}(z)$ that appear in (1.4) change very rapidly, but *smoothly*, from being exponentially small to being almost constant as ph z passes continuously through either of the values $\pm\pi$.

The proof of Theorem 1.1 was based on an integral representation of the function $U(a, a-b+1, z)$. By suitable expansion of this representation, the remainder term $R_n(a, b, z)$ was expressed as a double integral, from which (1.4) was generated by appropriate transformations and re-expansion. Now the expansion (1.1) can also be obtained by direct application of the asymptotic theory of linear differential equations; see, for example, [5, Chap. 7, § 10]. Two questions naturally arise. Can the expansion (1.4) also be derived by a purely differential-equation approach? If so, can a more general theory be constructed, indeed, a theory that would be applicable in cases where no suitable integral representation of the wanted solution exists? Again, the formal researches of Berry [2] indicate that the answers to both questions are in the affirmative. The purpose of the present paper is to provide a rigorous proof of Theorem 1.1 based on differential-equation theory. The analysis we shall develop serves as a valuable preliminary to the more general case, treated in [10]. Other relevant references are [4] and [11].

In proving Theorem 1.1 we shall concentrate on the result (1.5) for the central sector $|\text{ph } z|\leqq\pi$. The result (1.6) in the abutting sectors $\pi\leqq|\text{ph } z|\leqq\frac{5}{2}\pi-\delta$ may then be derived by application of well-known connection formulae for $U(a, a-b+1, z)$; compare [7, § 5].

2. Properties of the function $F_p(z)$. We collect here results pertaining to the function $F_p(z)$ that will be needed in the present paper and its sequel [10]. In each result p and z may be real or complex.

In terms of the generalized exponential integral $E_p(z)$ and the incomplete Gamma function $\Gamma(1-p, z)$, $F_p(z)$ is given by

$$F_p(z)=\frac{\Gamma(p)}{2\pi}\frac{E_p(z)}{z^{p-1}}=\frac{\Gamma(p)}{2\pi}\Gamma(1-p,z). \tag{2.1}$$

From the definition (1.7) we easily derive the following identities:

$$F_{p+1}(z)+F_p(z)=\frac{\Gamma(p)}{2\pi}\frac{e^{-z}}{z^p}, \tag{2.2}$$

$$F_{p+2}(z)+F_{p+1}(z)=\frac{p}{z}\{F_{p+1}(z)+F_p(z)\}, \tag{2.3}$$

$$F_p'(z)=-F_{p+1}(z)-F_p(z), \qquad \frac{d}{dz}\{e^zF_p(z)\}=-e^zF_{p+1}(z). \tag{2.4}$$

By rotation of the integration path we derive the following continuation formula in which k denotes any integer:

$$F_p(ze^{2k\pi i})=ie^{-kp\pi i}\frac{\sin(kp\pi)}{\sin(p\pi)}+e^{-2kp\pi i}F_p(z), \qquad z\neq 0. \tag{2.5}$$

When $|z|$ is large and p is fixed, or bounded, we find by application of Watson's lemma [5, Chap. 4, § 3],

$$F_p(z)\sim\frac{e^{-z}}{2\pi}\sum_{s=0}^{\infty}(-)^s\frac{\Gamma(p+s)}{z^{p+s}}, \qquad |\text{ph } z|\leqq\tfrac{3}{2}\pi-\delta\,(<\tfrac{3}{2}\pi). \tag{2.6}$$

Next, when $|p|$ and $|z|$ are both large,

$$F_p(z)=\tfrac{1}{2}i\,e^{-p\pi i}[\text{erfc}\,\{c(\theta)\sqrt{\tfrac{1}{2}|z|}\}+e^{-|z|\{c(\theta)\}^2/2}O(z^{-1/2})], \tag{2.7}$$

the O-term being uniform with respect to $\theta\equiv\text{ph } z$ in the interval $[-\pi+\delta, 3\pi-\delta]$, and also uniform with respect to bounded values of $|p-|z||$. Here $c(\theta)$ is the continuous branch of the function

$$c(\theta)=\{2\,e^{i\theta}+2i(\theta-\pi)+2\}^{1/2} \tag{2.8}$$

that is asymptotic to $\pi-\theta$ as $\theta\to\pi$. The approximation (2.7) is derived in [6], and a graph of $c(\theta)$ is also included in this reference.[1] As a consequence of these results, the corresponding approximation for the sector $-3\pi+\delta\leq\text{ph } z\leqq\pi-\delta$, and the well-known asymptotic estimates

$$\text{erfc}\,(\zeta)=O(e^{-\zeta^2}\zeta^{-1}), \qquad \zeta\to\infty \text{ in } |\text{ph } \zeta|\leqq\tfrac{3}{4}\pi-\delta,$$
$$\text{erfc}\,(\zeta)=2+O(e^{-\zeta^2}\zeta^{-1}), \qquad \zeta\to\infty \text{ in } |\text{ph } (-\zeta)|\leqq\tfrac{3}{4}\pi-\delta,$$

we derive the following uniform estimates, again valid when $|z|$ is large and $|p-|z||$ is bounded:

$$F_p(z)=e^{-z-|z|}O(1), \qquad |\text{ph } z|\leqq\pi, \tag{2.9}$$

$$F_p(z)=e^{-z-|z|}O(z^{-1/2}), \qquad |\text{ph } z|\leqq\pi-\delta, \tag{2.10}$$

$$F_p(z)=\pm i\,e^{\mp p\pi i}+e^{-z-|z|}O(1), \qquad \pi\leqq\pm\text{ph } z\leqq 3\pi-\delta, \tag{2.11}$$

$$F_p(z)=\pm i\,e^{\mp p\pi i}+e^{-z-|z|}O(z^{-1/2}), \qquad \pi+\delta\leqq\pm\text{ph } z\leqq 3\pi-\delta. \tag{2.12}$$

In (2.11) either all upper signs are taken, or all lower signs; (2.12) is similar.

[1] The result (2.7) is contained in the proof of Theorem 1 of [6], rather than in the actual statement of this theorem.

3. Matching of the remainder term by a series of F-functions. The function $U(a, a-b+1, z)$ satisfies the equation

$$Lw=0, \tag{3.1}$$

in which L denotes the differential operator

$$\frac{d^2}{dz^2}+\left(\frac{a-b+1}{z}-1\right)\frac{d}{dz}-\frac{a}{z}.$$

If n is an arbitrary nonnegative integer and we apply the operator L to the nth partial sum of the series (1.1), then we find that

$$L\left\{z^{-a}\sum_{s=0}^{n-1}(-)^s\frac{(a)_s(b)_s}{s!\,z^s}\right\}=(-)^{n-1}\frac{A_n}{z^{n+a+1}}, \tag{3.2}$$

where

$$A_n=\frac{(a)_n(b)_n}{(n-1)!}=\frac{\Gamma(n+a)\Gamma(n+b)}{\Gamma(a)\Gamma(b)(n-1)!}. \tag{3.3}$$

Hence from (3.1) it follows that the remainder term in (1.2) satisfies

$$LR_n(a,b,z)=(-)^n\frac{A_n}{z^{n+a+1}}. \tag{3.4}$$

Our objective is to show that $R_n(a, b, z)$ can be expanded in the form (1.4) in which m is an arbitrary nonnegative integer, and the new remainder $R_{m,n}(a, b, z)$ is estimated by (1.5) and (1.6). We shall approach this problem by constructing a finite series of F-functions which, when operated upon by L, will match the right-hand side of (3.4) except for an asymptotically small term.

Consider first $L\{e^zF_p(z)/z^q\}$, where p and q are arbitrary constants. With the aid of (2.4) we find that

$$\frac{d}{dz}\left\{\frac{e^zF_p(z)}{z^q}\right\}=-\frac{e^zF_{p+1}(z)}{z^q}-q\frac{e^zF_p(z)}{z^{q+1}},$$

and hence

$$\frac{d^2}{dz^2}\left\{\frac{e^zF_p(z)}{z^q}\right\}=\frac{e^zF_{p+2}(z)}{z^q}+2q\frac{e^zF_{p+1}(z)}{z^{q+1}}+q(q+1)\frac{e^zF_p(z)}{z^{q+2}}.$$

As a consequence,

$$L\left\{\frac{e^zF_p(z)}{z^q}\right\}=\frac{e^zF_{p+2}(z)}{z^q}+(2q-a+b-1+z)\frac{e^zF_{p+1}(z)}{z^{q+1}}$$
$$+\{q(q-a+b)+(q-a)z\}\frac{e^zF_p(z)}{z^{q+2}}.$$

The term in $F_{p+2}(z)$ can be eliminated with the aid of (2.3); thus

$$L\left\{\frac{e^zF_p(z)}{z^q}\right\}=(p+2q-a+b-1)\frac{e^zF_{p+1}(z)}{z^{q+1}}+\{q(q-a+b)+(p+q-a)z\}\frac{e^zF_p(z)}{z^{q+2}}.$$

This equation may be rearranged in the form

$$L\left\{\frac{e^zF_p(z)}{z^q}\right\}=(p+q-a)\left\{\frac{e^zF_{p+1}(z)}{z^{q+1}}+\frac{e^zF_p(z)}{z^{q+1}}\right\}+(q+b-1)\frac{e^zF_{p+1}(z)}{z^{q+1}}$$
$$+q(q-a+b)\frac{e^zF_p(z)}{z^{q+2}}.$$

Then by use of (2.2) we arrive at

$$L\left\{\frac{e^zF_p(z)}{z^q}\right\}=\frac{p+q-a}{2\pi}\frac{\Gamma(p)}{z^{p+q+1}}+(q+b-1)\frac{e^zF_{p+1}(z)}{z^{q+1}}+q(q-a+b)\frac{e^zF_p(z)}{z^{q+2}}. \tag{3.5}$$

In order to make the asymptotic behaviour of the first term on the right-hand side of (3.5) match that of the right-hand side of (3.4), we must first arrange that the powers of z agree; thus

$$p+q=n+a. \tag{3.6}$$

From (3.3) and Stirling's formula, we derive

$$A_n\sim(\text{constant})\times e^{-n}n^{n+a+b-(1/2)},\qquad n\to\infty.$$

Also, with $q=n+a-p$,

$$(p+q-a)\Gamma(p)\sim(\text{constant})\times n\,e^{-p}p^{p-(1/2)},\qquad p\to\infty.$$

Evidently, apart from the constant factors, the asymptotic behaviour of these two expressions will be the same if we choose $p=n+a+b-1$ and, correspondingly, $q=1-b$. With these choices (3.5) becomes

$$L\left\{\frac{e^zF_{n+a+b-1}(z)}{z^{1-b}}\right\}=\frac{n}{2\pi}\frac{\Gamma(n+a+b-1)}{z^{n+a+1}}+(1-a)(1-b)\frac{e^zF_{n+a+b-1}(z)}{z^{3-b}}. \tag{3.7}$$

The second term on the right-hand side of the last equation is to be regarded as a residual, and to drive it into another residual, but with a higher power of z in the denominator, we consider (3.5) with q increased by unity, to $2-b$. At the same time, to preserve the asymptotic form of the other contribution, we decrease p by unity, to $n+a+b-2$; compare (3.6). Thus we have

$$L\left\{\frac{e^zF_{n+a+b-2}(z)}{z^{2-b}}\right\}=\frac{n}{2\pi}\frac{\Gamma(n+a+b-2)}{z^{n+a+1}}+\frac{e^zF_{n+a+b-1}(z)}{z^{3-b}}$$
$$+(2-a)(2-b)\frac{e^zF_{n+a+b-2}(z)}{z^{4-b}}.$$

Then by eliminating the final term from (3.7), we arrive at

$$\begin{aligned}&L\left\{\frac{e^zF_{n+a+b-1}(z)}{z^{1-b}}-(1-a)(1-b)\frac{e^zF_{n+a+b-2}(z)}{z^{2-b}}\right\}\\&=\frac{n}{2\pi z^{n+a+1}}\{\Gamma(n+a+b-1)-(1-a)(1-b)\Gamma(n+a+b-2)\}\\&\quad-(1-a)_2(1-b)_2\frac{e^zF_{n+a+b-2}(z)}{z^{4-b}}.\end{aligned} \tag{3.8}$$

The procedure may be repeated. The last term of the right member of (3.8) is now regarded as the residual, and we drive it into another residual with yet a higher power of z in the denominator by use of (3.5) with $p=n+a+b-3$ and $q=3-b$. Continuing in this manner we readily establish, by induction, the general result

$$\begin{aligned}&L\left\{\sum_{s=0}^{m-1}(-)^s\frac{(1-a)_s(1-b)_s}{s!}\frac{e^zF_{n+a+b-s-1}(z)}{z^{s-b+1}}\right\}\\&=\frac{nB_{m,n}}{2\pi z^{n+a+1}}+(-)^{m-1}\frac{(1-a)_m(1-b)_m}{(m-1)!}\frac{e^zF_{n+a+b-m}(z)}{z^{m-b+2}},\end{aligned} \tag{3.9}$$

in which m is arbitrary, and

$$B_{m,n}=\sum_{s=0}^{m-1}(-)^s\frac{(1-a)_s(1-b)_s}{s!}\Gamma(n+a+b-s-1). \tag{3.10}$$

On comparing (3.4) and (3.9) we perceive that the desired matching is representable in the form

$$L\left\{R_n(a,b,z)+(-)^{n-1}\frac{2\pi A_n}{nB_{m,n}}\sum_{s=0}^{m-1}(-)^s\frac{(1-a)_s(1-b)_s}{s!}\frac{e^zF_{n+a+b-s-1}(z)}{z^{s-b+1}}\right\}$$
$$=(-)^{m+n}\frac{2\pi A_n}{nB_{m,n}}\frac{(1-a)_m(1-b)_m}{(m-1)!}\frac{e^zF_{n+a+b-m}(z)}{z^{m-b+2}}.$$

In § 4 we shall, in effect, invert the operator L in the last equation in order to estimate asymptotically the content of the braces in the left member. However, by comparison with the known result (1.4) we suspect that a simplifying approximation can be made for the ratio $A_n/(nB_{m,n})$. This is indeed the case, and the required result is as follows.

LEMMA 3.1. *Let a and b be fixed (or bounded) complex numbers and m be a fixed nonnegative integer. Then as $n\to\infty$*

$$\frac{\Gamma(n+a)\Gamma(n+b)}{\Gamma(n+1)}=\sum_{s=0}^{m-1}(-)^s\frac{(1-a)_s(1-b)_s}{s!}\Gamma(n+a+b-s-1)$$
$$+(1-a)_m(1-b)_m\Gamma(n+a+b-m-1)O(1).$$

This result is proved in [12, Appendix]. See also [8], [9].

4. Estimation of $R_{m,n}(a, b, z)$. Throughout this section it will be understood that the parameters a and b are fixed (or bounded).

Let us return to (3.4) and (3.9). By combination, we obtain

$$LR_n(a,b,z)-L\left\{(-)^n\frac{2\pi}{\Gamma(a)\Gamma(b)}\sum_{s=0}^{m-1}(-)^s\frac{(1-a)_s(1-b)_s}{s!}\frac{e^zF_{n+a+b-s-1}(z)}{z^{s-b+1}}\right\}$$
$$=(-)^n\left\{A_n-\frac{n}{\Gamma(a)\Gamma(b)}B_{m,n}\right\}\frac{1}{z^{n+a+1}}$$
$$+(-)^{m+n}\frac{2\pi(1-a)_m(1-b)_m}{(m-1)!\,\Gamma(a)\Gamma(b)}\frac{e^zF_{n+a+b-m}(z)}{z^{m-b+2}}.$$

Hence, with $R_{m,n}(a, b, z)$ defined as in (1.4), we have

$$L\{e^zz^{b-1}R_{m,n}(a,b,z)\}=\frac{C_{m,n}}{z^{n+a+1}}+\frac{(-)^m}{(m-1)!}\frac{e^zF_{n+a+b-m}(z)}{z^{m-b+2}}, \tag{4.1}$$

where

$$C_{m,n}=\frac{\Gamma(a)\Gamma(b)}{2\pi(1-a)_m(1-b)_m}\left\{A_n-\frac{n}{\Gamma(a)\Gamma(b)}B_{m,n}\right\}. \tag{4.2}$$

And as a consequence of (3.3), (3.10), and Lemma 3.1 we note that as $n\to\infty$, with m fixed,

$$C_{m,n}=n\Gamma(n+a+b-m-1)O(1)=n^{n+a+b-m-(1/2)}e^{-n}O(1). \tag{4.3}$$

In order to invert the operator L in (4.1) we need two linearly independent solutions of the equation $Lw=0$. One of these has to be the solution that is recessive at infinity in the sector $|\text{ph}\, z|<\frac{1}{2}\pi$, that is, $U(a, a-b+1, z)$. As a second solution we select $V(a, a-b+1, z)$, the solution that is recessive at infinity in the sector $|\text{ph}\,(-z)|<\frac{1}{2}\pi$. For brevity, we shall denote these solutions by $U(z)$ and $V(z)$, respectively. Then their defining properties are expressed by

$$(4.4)\qquad U(z)=z^{-a}\{1+O(z^{-1})\},\qquad z\to\infty \text{ in } |\text{ph}\, z|\leqq\tfrac{3}{2}\pi-\delta;$$

$$(4.5)\qquad V(z)=e^{z}(-z)^{b-1}\{1+O(z^{-1})\},\qquad z\to\infty \text{ in } |\text{ph}\,(-z)|\leqq\tfrac{3}{2}\pi-\delta;$$

compare [5, Chap. 7, § 10.1].

Until the end of this section, we shall restrict $0\leqq\text{ph}\, z\leqq\pi$. The branch of z^{-a} in (4.4) is then the principal value, and the choice of branch of $(-z)^{b-1}$ in (4.5) is given by

$$(4.6)\qquad V(z)=e^{\pi i(1-b)}e^{z}z^{b-1}\{1+O(z^{-1})\},\qquad z\to\infty \text{ in } 0\leqq\text{ph}\, z\leqq\pi,$$

where z^{b-1} has its principal value. The Wronskian of the two solutions is

$$(4.7)\qquad \mathscr{W}\{U(z), V(z)\}=e^{\pi i(1-b)}e^{z}z^{b-a-1}.$$

By the method of variation of parameters, using (4.7), we find that one solution of (4.1) is given by

$$(4.8)\qquad R_{m,n}(a, b, z)=R^{(1)}_{m,n}(a, b, z)+R^{(2)}_{m,n}(a, b, z),$$

where

$$(4.9)\qquad e^{z}z^{b-1}R^{(1)}_{m,n}(a, b, z)=C_{m,n}\int_{z}^{\infty}\frac{K(z, t)}{t^{n+a+1}}\,dt,$$

$$(4.10)\qquad e^{z}z^{b-1}R^{(2)}_{m,n}(a, b, z)=\frac{(-)^{m}}{(m-1)!}\int_{z}^{\infty}\frac{K(z, t)}{t^{m-b+2}}\,e^{t}F_{n+a+b-m}(t)\,dt,$$

and

$$(4.11)\qquad K(z, t)=e^{\pi i(b-1)}\,e^{-t}t^{a-b+1}\{U(z)V(t)-V(z)U(t)\}.$$

From (2.6), (4.4), and (4.6) we may verify that as $z\to\infty$, the right members of (4.9) and (4.10) are $O(z^{-n-a})$ and $O(z^{-n-a-1})$, respectively. Hence with $R_{m,n}(a, b, z)$ defined by (4.8) it follows that $e^{z}z^{b-1}R_{m,n}(a, b, z)$ is $O(z^{-n-a})$. Again, with the aid also of (2.6) we see that the right-hand side of (1.4) becomes $O(z^{-n-a})$. This is precisely what is required by (1.2), and since there is only one solution of (3.4) that is $O(z^{-n-a})$ as $z\to\infty$, it follows that (4.8) *is* the correct choice of solution of (4.1) when $R_{m,n}(a, b, z)$ is defined by (1.2) and (1.4).

In order to estimate $R_{m,n}(a, b, z)$ from (4.8), (4.9), and (4.10), we need a uniform estimate of $K(z, t)$ when $|z|$ is large. This estimate will depend on the integration paths used in (4.9) and (4.10). We shall employ the same path for each integral, and our first restrictions are that $0\leqq\text{ph}\, t\leqq\pi$ and $|t|\geqq|z|$ everywhere on this path. Then from (4.4) and (4.6) we conclude that

$$(4.12)\qquad U(z)V(t)=z^{-a}\,e^{t}t^{b-1}O(1)$$

when $|z|$ is large in the sector $0\leqq\text{ph}\, z\leqq\pi$, uniformly for t on the path. As is customary in proofs of this nature, we now try to arrange that the same estimate applies to the other constituent $V(z)U(t)$ in the kernel $K(z, t)$. With this in view we prescribe the integration path in (4.9) and (4.10) to consist of: (i) a circular arc from $t=z$ to $t=|z|$, centered at the origin and described in the positive sense; (ii) the real axis from $t=|z|$

to $t=\infty$. This path is illustrated in Fig. 4.1. If z is real and positive, the circular arc is absent.

LEMMA 4.1. *Assume that*

$$|z| \geqq \tfrac{1}{2}\pi|a+b-1| \quad \textit{and} \quad 0 \leqq \operatorname{ph} z \leqq \pi.$$

Then on the integration path depicted in Fig. 4.1.

$$\left| e^{z-t}\left(\frac{z}{t}\right)^{a+b-1} \right| \leqq \exp\left\{ \frac{\pi^2 |\operatorname{Im}(a+b)|^2}{8|z|} \right\},$$

provided that $(z/t)^{a+b-1}$ *has its principal value.*

This result is proved in the Appendix. We note, in passing, that the bound is unity when $a+b$ is real, and is $1+O(z^{-1})$ in other circumstances.

Corresponding to (4.12), we have

$$V(z)U(t) = e^z z^{b-1} t^{-a} O(1),$$

and hence, by Lemma 4.1,

$$V(z)U(t) = z^{-a} e^t t^{b-1} O(1). \tag{4.13}$$

On substituting into (4.11) by means of (4.12) and (4.13) we obtain the desired estimate

$$K(z,t) = z^{-a} t^a O(1), \tag{4.14}$$

uniformly valid when t lies in the path depicted in Fig. 4.1 and $|z|$ is large.

Next, on substituting into (4.9) by means of (4.14), we obtain

$$e^z z^{b-1} R^{(1)}_{m,n}(a,b,z) = C_{m,n} z^{-a} O(1) \int_z^\infty \left| \frac{dt}{t^{n+1}} \right|.$$

On the chosen path

$$\int_z^\infty \left| \frac{dt}{t^{n+1}} \right| \leqq \frac{\pi}{|z|^n} + \frac{1}{n|z|^n}. \tag{4.15}$$

Hence we have

$$R^{(1)}_{m,n}(a,b,z) = C_{m,n} \frac{e^{-z}}{z^{n+a+b-1}} \left(\pi + \frac{1}{n} \right) O(1). \tag{4.16}$$

So far in the analysis we have not assumed any relationship between $|z|$ and n. It is at this stage that we introduce the condition (1.3) in which a, b, and α are bounded. Since $n \to \infty$ as $|z| \to \infty$, we may substitute for $C_{m,n}$ in (4.16) by means of (4.3). We obtain

$$R^{(1)}_{m,n}(a,b,z) = n^{n+a+b-m-(1/2)} e^{-n} \frac{e^{-z}}{z^{n+a+b-1}} O(1) = e^{-z-|z|} z^{-m+(1/2)} O(1). \tag{4.17}$$

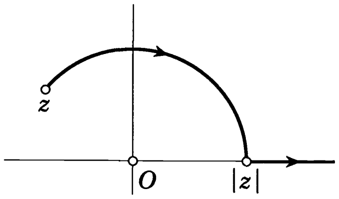


FIG. 4.1. *t-plane. Path for the integrals* (4.9) *and* (4.10). *Sector* $0 \leqq \operatorname{ph} z \leqq \pi$.

The corresponding estimation of $R^{(2)}_{m,n}(a, b, z)$ from (4.10) is more complicated. Consider first the circular arc from z to $|z|$. On this part of the path $|n+a+b-m-|t||$ is bounded; compare (1.3). Hence from (2.9) we obtain

$$F_{n+a+b-m}(t)=e^{-t-|t|}O(1).$$

Using this result and (4.14) we derive

$$\int_z^{|z|}\frac{K(z,t)}{t^{m-b+2}}e^tF_{n+a+b-m}(t)\,dt=z^{-a}O(1)\int_z^{|z|}\frac{e^{-|t|}|dt|}{|t^{m-a-b+2}|}=\frac{e^{-|z|}}{z^{m-b+1}}O(1). \tag{4.18}$$

Now consider the straight line segment from $|z|$ to ∞. On referring to (1.7) and (4.14) we perceive that

$$\begin{aligned}&\int_{|z|}^{\infty}\frac{K(z,t)}{t^{m-b+2}}e^tF_{n+a+b-m}(t)\,dt\\&\quad=O(z^{-a})\int_{|z|}^{\infty}\frac{dt}{t^{m-\operatorname{Re}a-\operatorname{Re}b+2}}\int_0^{\infty}\frac{e^{-t\tau}\tau^{n+\operatorname{Re}a+\operatorname{Re}b-m-1}}{1+\tau}\,d\tau.\end{aligned}$$

Provided that $m>\operatorname{Re}a+\operatorname{Re}b-1$, the integration order may be interchanged [3, Thm. 1]; thus

$$\begin{aligned}&\int_{|z|}^{\infty}\frac{K(z,t)}{t^{m-b+2}}e^tF_{n+a+b-m}(t)\,dt\\&\quad=O(z^{-a})\int_0^{\infty}\frac{\tau^{n+\operatorname{Re}a+\operatorname{Re}b-m-1}}{1+\tau}\,d\tau\int_{|z|}^{\infty}\frac{e^{-t\tau}}{t^{m-\operatorname{Re}a-\operatorname{Re}b+2}}\,dt.\end{aligned}$$

In terms of the incomplete Gamma function the inner integral on the right-hand side of the last equation is

$$\tau^{m-\operatorname{Re}a-\operatorname{Re}b+1}\Gamma(\operatorname{Re}a+\operatorname{Re}b-m-1,|z|\tau).$$

Since we are assuming that $m>\operatorname{Re}a+\operatorname{Re}b-1$, this quantity is bounded by

$$e^{-|z|\tau}|z|^{\operatorname{Re}a+\operatorname{Re}b-m-2}\tau^{-1};$$

compare [5, Chap. 3, eq. (1.05)]. Hence we see that

$$\begin{aligned}\int_{|z|}^{\infty}\frac{K(z,t)}{t^{m-b+2}}e^tF_{n+a+b-m}(t)\,dt&=\frac{O(1)}{z^{m-b+2}}\int_0^{\infty}\frac{e^{-|z|\tau}\tau^{n+\operatorname{Re}a+\operatorname{Re}b-m-2}}{1+\tau}\,d\tau\\&=e^{|z|}z^{b-m-2}F_{n+\operatorname{Re}a+\operatorname{Re}b-m-1}(|z|)O(1)=e^{-|z|}z^{b-m-2}O(1);\end{aligned} \tag{4.19}$$

compare (1.7) and (2.9).

We may now substitute into (4.10) by means of (4.18) and (4.19) to obtain

$$R^{(2)}_{m,n}(a,b,z)=e^{-z-|z|}z^{-m}O(1). \tag{4.20}$$

Then from (4.8), (4.17), and (4.20) we arrive at

$$R_{m,n}(a,b,z)=e^{-z-|z|}O(z^{-m+(1/2)}) \tag{4.21}$$

as $z\to\infty$ in the sector $0\leqq\operatorname{ph}z\leqq\pi$. A similar proof applies when $-\pi\leqq\operatorname{ph}z\leqq 0$.

To complete the proof of (1.5) we need to show that: (i) the order term in (4.21) can be strengthened to $O(z^{-m})$; (ii) the condition $m>\operatorname{Re}a+\operatorname{Re}b-1$ can be removed. Both steps are easily achieved simply by increasing m in (1.4) and using the estimate

$$F_{n+a+b-s-1}(z)=e^{-z-|z|}O(1)$$

as necessary for $s\leqq m-1$; compare, again, (2.9).

5. Conclusions. We have developed a new form of asymptotic analysis for the study of solutions of linear differential equations in order to establish an exponentially-improved asymptotic expansion of the confluent hypergeometric function $U(a, a-b+1, z)$ in the neighbourhood of the irregular singularity at $z=\infty$. This expansion was derived previously from an integral representation of $U(a, a-b+1, z)$. Our proof applies directly to the sector $|\text{ph } z| \leqq \pi$, which is the region of maximum exponential improvement. The corresponding results for the sectors $\pi \leqq |\text{ph } z| \leqq \frac{5}{2}\pi - \delta$ can be obtained by application of connection formulae.

The integral-representation approach also led to strict and realistic error bounds for the exponentially-improved asymptotic expansion of $U(a, a-b+1, z)$ in the sector $|\text{ph } z| \leqq \pi$. It may be feasible to obtain similar error bounds via our differential-equation approach, but this has not been explored in the present paper.

Appendix. Proof of Lemma 4.1. Write

$$\Phi(t) = \text{Re}\,\{t + (a+b-1)\ln t\},$$

where the logarithm assumes its principal value. Then on the path depicted in Fig. 4.1 we have

$$\left| e^{z-t}\left(\frac{z}{t}\right)^{a+b-1} \right| = e^{\Phi(z)-\Phi(t)}.$$

Let us consider the behaviour of $\Phi(t)$ at t passes along the contour $ABCD$ shown in Fig. A.1. Here ABC is the semicircle $t = \rho\, e^{i\chi}$, where $\rho = |z|$ and $\pi \geqq \chi \geqq 0$; CD is the segment of the real axis from $t=\rho$ to $t=\infty$. We shall also write

$$a+b-1 = \kappa_{\text{R}} + i\kappa_{\text{I}},$$

where κ_{R} and κ_{I} are real.

On CD we have

$$\Phi(t) = t + \kappa_{\text{R}} \ln t, \qquad \Phi'(t) = 1 + (\kappa_{\text{R}}/t).$$

Hence $\Phi(t)$ is nondecreasing, provided that $t \geqq -\kappa_{\text{R}}$ in the case $\kappa_{\text{R}} < 0$.

Now suppose that t lies on ABC and $\kappa_{\text{I}} \geqq 0$. Then

$$\Phi(t) = \rho\cos\chi + \kappa_{\text{R}}\ln\rho - \kappa_{\text{I}}\chi, \qquad \frac{d\Phi}{d\chi} = -\rho\sin\chi - \kappa_{\text{I}}.$$

Hence $\Phi(t)$ is nondecreasing.

Lastly, suppose that t lies on ABC and $\kappa_{\text{I}} = -\kappa'_{\text{I}}$, where $\kappa'_{\text{I}} > 0$. Then

$$\Phi(t) = \rho\cos\chi + \kappa_{\text{R}}\ln\rho + \kappa'_{\text{I}}\chi, \quad \frac{d\Phi}{d\chi} = -\rho\sin\chi + \kappa'_{\text{I}}, \quad \frac{d^2\Phi}{d\chi^2} = -\rho\cos\chi.$$

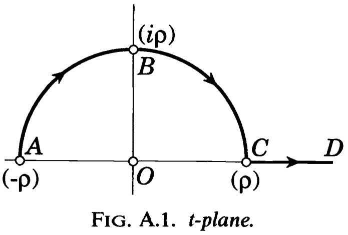


FIG. A.1. *t-plane.*

Suppose also that $\rho > \kappa_{\mathrm{I}}'$. Then $d\Phi/d\chi$ vanishes at $\chi = \chi_0$ and $\chi = \pi - \chi_0$, where $\chi_0 = \sin^{-1}(\kappa_{\mathrm{I}}'/\rho)$. The graph of $\Phi(t)$ plotted against χ is illustrated in Fig. A.2.

At C, $t = \rho$, $\chi = 0$, and $\Phi(t) = \Phi(\rho) = \rho + \kappa_{\mathrm{R}} \ln \rho$.

At $t = \rho\, e^{i\chi_0}$, $\Phi(t)$ is at a maximum, and we find that

$$\Phi(\rho\, e^{i\chi_0}) - \Phi(\rho) = \rho \cos \chi_0 - \rho + \kappa_{\mathrm{I}}'\chi_0 = h,$$

say. Similarly, at $t = \rho\, e^{i(\pi - \chi_0)}$, $\Phi(t)$ is at a minimum, and we find that

$$\Phi(\rho\, \mathrm{e}^{i\pi}) - \Phi(\rho\, e^{i(\pi-\chi_0)}) = h,$$

with the same h. Also, because $\chi_0 = \sin^{-1}(\kappa_{\mathrm{I}}'/\rho)$, we have

$$h = \rho(\cos \chi_0 - 1 + \chi_0 \sin \chi_0).$$

Since $0 \leqq \chi_0 \leqq \frac{1}{2}\pi$, we have by an elementary inequality $0 \leqq h \leqq \frac{1}{2}\rho\chi_0^2$, and hence by Jordan's inequality

$$0 \leqq h \leqq \tfrac{1}{2}\rho\left(\frac{\pi\kappa_{\mathrm{I}}'}{2\rho}\right)^2 = \frac{\pi^2\kappa_{\mathrm{I}}'^2}{8\rho}.$$

We require one further condition, given by $\rho \geqq \frac{1}{2}\pi\kappa_{\mathrm{I}}'$. The purpose of this condition is to ensure that $\Phi(\rho) \geqq \Phi(\rho\, e^{i\pi})$ and, in consequence, that the point C in Fig. A.2 does not lie below A.[2]

On combining the foregoing results we see that we can assert that if z and t are any two points on the path depicted in Fig. A.1, t being to the right of z, then in all cases

$$\Phi(z) - \Phi(t) \leqq h \leqq \frac{\pi^2\kappa_{\mathrm{I}}^2}{8\rho},$$

provided that $|z| \geqq -\kappa_{\mathrm{R}}$ when κ_{R} is negative, and $|z| \geqq -\frac{1}{2}\pi\kappa_{\mathrm{I}}$ when κ_{I} is negative. The last two conditions are certainly fulfilled when $|z| \geqq \frac{1}{2}\pi|\kappa_{\mathrm{R}} + i\kappa_{\mathrm{I}}| = \frac{1}{2}\pi|a + b - 1|$, and Lemma 4.1 follows.

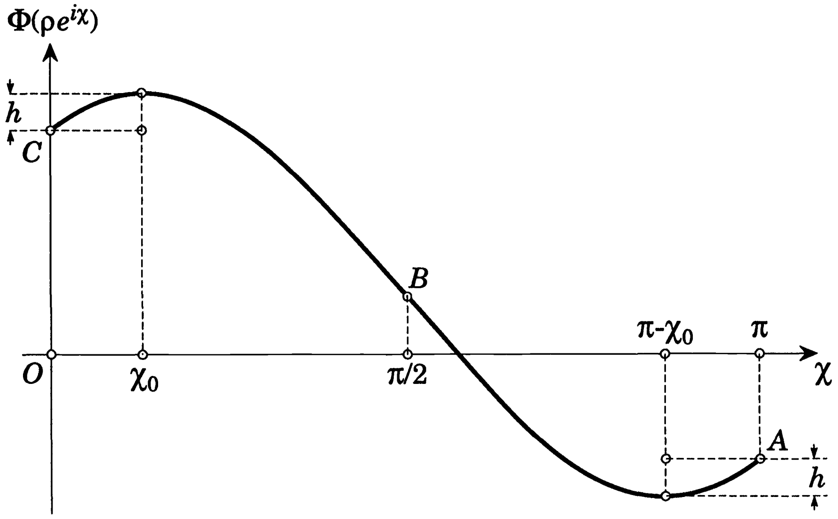


FIG. A.2. *Graph of $\Phi(\rho e^{i\chi})$ when κ_{I} is negative.*

[2] Actually, it would suffice for C to be no lower than the local minimum at $\chi = \pi - \chi_0$.

Acknowledgment. The writer is pleased to acknowledge helpful discussions with R.B. Paris and A. D. Wood.

REFERENCES

[1] M. V. BERRY, *Uniform asymptotic smoothing of Stokes's discontinuities*, Proc. Roy. Soc. London Ser. A, 422 (1989), pp. 7-21.

[2] ———, *Waves near Stokes lines*, Proc. Roy. Soc. London Ser. A, 427 (1990), pp. 265-280.

[3] E. R. LOVE, *Changing the order of integration*, J. Austral. Math. Soc., 11 (1970), pp. 421-432.

[4] J. B. MCLEOD, *Smoothing of Stokes discontinuities*, Proc. Roy. Soc. London Ser. A, 437 (1992), pp. 343-354.

[5] F. W. J. OLVER, Asymptotics and Special Functions, Academic Press, New York, 1974.

[6] ———, *Uniform, exponentially improved, asymptotic expansions for the generalized exponential integral*, SIAM J. Math. Anal., 22 (1991), pp. 1460-1474.

[7] ———, *Uniform, exponentially improved, asymptotic expansions for the confluent hypergeometric function and other integral transforms*, SIAM J. Math. Anal., 22 (1991), pp. 1475-1489.

[8] ———, *Asymptotic expansions of the coefficients in asymptotic series solutions of linear differential equations*, manuscript.

[9] ———, *On an asymptotic expansion of a ratio of Gamma functions*, manuscript.

[10] ———, *Exponentially-improved asymptotic solutions of ordinary differential equations* II: *Irregular singularities of rank one*, manuscript.

[11] R. B. PARIS, *Smoothing of the Stokes phenomenon for high-order differential equations*, Proc. Roy. Soc. London Ser. A, 436 (1992), pp. 165-186.

[12] ———, *Smoothing of the Stokes phenomenon using Mellin-Barnes integrals*, J. Comput. Appl. Math., 41 (1992), pp. 117-133.

SIAM J. MATH. ANAL.
Vol. 24, No. 3, pp. 768-782, May 1993

014

A GENERALIZATION OF LAGUERRE POLYNOMIALS*

R. KOEKOEK† AND H. G. MEIJER†

Abstract. The authors study orthogonal polynomials on $[0,+\infty)$ with respect to an inner product involving derivatives that cannot be derived from a weight function. These polynomials can be written as a ${}_3F_3$ hypergeometric series and they satisfy a second-order differential equation and a five term recurrence relation. At most one zero of each polynomial is located outside $(0,+\infty)$, the interior of the interval of orthogonality. As a special case Koornwinder's Laguerre polynomials $\{L_n^{\alpha,M}(x)\}_{n=0}^{+\infty}$ are included.

Key words. orthogonal polynomials, Laguerre polynomials, inner product

AMS subject classifications. 33A65, 33C45

1. Introduction. In [8] and [9] H. L. Krall introduced some generalizations of classical orthogonal polynomials which are orthogonal with respect to a weight function consisting of the classical weight function together with a delta function at the endpoint(s) of the interval of orthogonality. These polynomials were described in more detail by A. M. Krall in [7]. In [6] Koornwinder generalized this and computed the polynomials which are orthogonal on the interval $[-1,+1]$ with respect to a weight function of the form $(1-x)^\alpha(1+x)^\beta+M\cdot\delta(x+1)+N\cdot\delta(x-1)$. These polynomials are generalizations of the classical Jacobi polynomials. In [1] and [2] Bavinck and Meijer studied further generalizations of these polynomials in the ultraspherical case $(\alpha=\beta)$; they computed the polynomials $\{S_n^{\alpha,M,N}(x)\}_{n=0}^{+\infty}$, which are orthogonal on the interval $[-1,+1]$ with respect to the inner product

$$\langle p,q\rangle=\frac{\Gamma(2\alpha+2)}{2^{2\alpha+1}\cdot\Gamma^2(\alpha+1)}\cdot\int_{-1}^{+1}(1-x^2)^\alpha\cdot p(x)q(x)\,dx+M\cdot[p(-1)q(-1)+p(1)q(1)]$$
$$+N\cdot[p'(-1)q'(-1)+p'(1)q'(1)].$$

As a limiting case Koornwinder found the polynomials $\{L_n^{\alpha,M}(x)\}_{n=0}^{+\infty}$, which are generalizations of the classical Laguerre polynomials $\{L_n^{(\alpha)}(x)\}_{n=0}^{+\infty}$ orthogonal on the interval $[0,+\infty)$ with respect to the weight function $(1/\Gamma(\alpha+1))\cdot x^\alpha e^{-x}+M\cdot\delta(x)$. These polynomials were described in detail in [4].

In the present paper we consider the polynomials $\{L_n^{\alpha,M,N}(x)\}_{n=0}^{+\infty}$, which are orthogonal with respect to the inner product:

$$(1.1)\quad \langle p,q\rangle=\frac{1}{\Gamma(\alpha+1)}\cdot\int_0^{+\infty}x^\alpha e^{-x}\cdot p(x)q(x)\,dx+M\cdot p(0)q(0)+N\cdot p'(0)q'(0),$$

where $\alpha>-1$, $M\geqq 0$, and $N\geqq 0$.

We will show that the polynomials $\{L_n^{\alpha,M,N}(x)\}_{n=0}^{+\infty}$ can be defined by

$$(1.2)\qquad L_n^{\alpha,M,N}(x)=A_0\cdot L_n^{(\alpha)}(x)+A_1\cdot\frac{d}{dx}L_n^{(\alpha)}(x)+A_2\cdot\frac{d^2}{dx^2}L_n^{(\alpha)}(x),$$

* Received by the editors July 5, 1988; accepted for publication (in revised form) August 31, 1992.

† Department of Mathematics and Informatics, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, the Netherlands.

where

$$(1.3)\begin{cases} A_0=1+M\cdot\binom{n+\alpha}{n-1}+\frac{n(\alpha+2)-(\alpha+1)}{(\alpha+1)(\alpha+3)}\cdot N\cdot\binom{n+\alpha}{n-2}+\frac{MN}{(\alpha+1)(\alpha+2)} \\ \qquad\cdot\binom{n+\alpha}{n-1}\binom{n+\alpha+1}{n-2}, \\ A_1=M\cdot\binom{n+\alpha}{n}+\frac{(n-1)}{(\alpha+1)}\cdot N\cdot\binom{n+\alpha}{n-1}+\frac{2MN}{(\alpha+1)^2}\cdot\binom{n+\alpha}{n}\binom{n+\alpha+1}{n-2}, \\ A_2=\frac{N}{(\alpha+1)}\cdot\binom{n+\alpha}{n-1}+\frac{MN}{(\alpha+1)^2}\cdot\binom{n+\alpha}{n}\binom{n+\alpha+1}{n-1}. \end{cases}$$

For $N=0$ we have

$$L_n^{\alpha,M,0}(x)=L_n^{\alpha,M}(x)=\left[1+M\cdot\binom{n+\alpha}{n-1}\right]\cdot L_n^{(\alpha)}(x)+M\cdot\binom{n+\alpha}{n}\cdot\frac{d}{dx}L_n^{(\alpha)}(x),$$

Koornwinder's polynomials, and for $M=N=0$ we have

$$L_n^{\alpha,0,0}(x)=L_n^{(\alpha)}(x),$$

the classical Laguerre polynomials.

In [5] we studied further generalizations of these polynomials $\{L_n^{\alpha,M_0,M_1,\ldots,M_k}(x)\}_{n=0}^{+\infty}$, which are orthogonal on the interval $[0,+\infty)$ with respect to an inner product involving higher derivatives:

$$\langle p,q\rangle=\frac{1}{\Gamma(\alpha+1)}\cdot\int_0^{+\infty}x^\alpha e^{-x}\cdot p(x)q(x)\,dx+\sum_{\nu=0}^{K}M_\nu\cdot p^{(\nu)}(0)q^{(\nu)}(0).$$

For $N>0$ the inner product (1.1) cannot be derived from a weight function, since $\langle 1,x^2\rangle\neq\langle x,x\rangle$. Many of the well-known properties of orthogonal polynomials can be proved by using the orthogonality with respect to a weight function (see [3] and [10]). So we may not expect these new polynomials to satisfy a three term recurrence relation and to have real simple zeros which lie in the interior of the interval of orthogonality.

In this paper we investigate some properties of the polynomials $\{L_n^{\alpha,M,N}(x)\}_{n=0}^{+\infty}$. In this special case we have an explicit representation which allows us to find results concerning the zeros, for instance. For general results on the polynomials $\{L_n^{\alpha,M_0,M_1,\cdots,M_k}(x)\}_{n=0}^{+\infty}$, the reader is referred to [5].

2. The classical Laguerre polynomials. First we summarize the properties of the classical Laguerre polynomials $\{L_n^{(\alpha)}(x)\}_{n=0}^{+\infty}$ we will use. For details the reader is referred to [3] and [10].

Let $\alpha>-1$. The classical Laguerre polynomials $\{L_n^{(\alpha)}(x)\}_{n=0}^{+\infty}$ are orthogonal polynomials on the interval $[0,+\infty)$ with respect to the weight function $x^\alpha e^{-x}$ and with the normalization

$$(2.1)\qquad L_n^{(\alpha)}(0)=\binom{n+\alpha}{n}.$$

They can be defined by Rodrigues' formula

$$(2.2)\qquad L_n^{(\alpha)}(x)=\frac{1}{n!}\cdot x^{-\alpha}e^x\cdot\frac{d^n}{dx^n}[e^{-x}\cdot x^{n+\alpha}],\qquad n=0,1,2,3,\ldots.$$

The polynomials $\{L_n^{(\alpha)}(x)\}_{n=0}^{+\infty}$ satisfy the Laguerre differential equation

$$x \cdot \frac{d^2}{dx^2} L_n^{(\alpha)}(x) + (\alpha+1-x) \cdot \frac{d}{dx} L_n^{(\alpha)}(x) + n \cdot L_n^{(\alpha)}(x) = 0, \tag{2.3}$$

and a three term recurrence relation

$$\begin{gathered}(n+1) \cdot L_{n+1}^{(\alpha)}(x) + (x-2n-\alpha-1) \\ \cdot L_n^{(\alpha)}(x) + (n+\alpha) \cdot L_{n-1}^{(\alpha)}(x) = 0, \qquad n = 1, 2, 3, 4, \ldots, \\ L_0^{(\alpha)}(x) = 1 \quad \text{and} \quad L_1^{(\alpha)}(x) = \alpha + 1 - x.\end{gathered}$$

We have a simple differentiation formula:

$$\frac{d}{dx} L_n^{(\alpha)}(x) = -L_{n-1}^{(\alpha+1)}(x), \qquad n = 1, 2, 3, 4, \ldots \tag{2.4}$$

and a representation as a hypergeometric series

$$L_n^{(\alpha)}(x) = \binom{n+\alpha}{n} \cdot {}_1F_1(-n; \alpha+1; x), \qquad n = 0, 1, 2, 3, \ldots. \tag{2.5}$$

3. The orthogonality. In this section we prove the orthogonality of the polynomials $\{L_n^{\alpha,M,N}(x)\}_{n=0}^{+\infty}$ defined by (1.2) and (1.3) with respect to the inner product (1.1).

We will prove that $\langle p, L_n^{\alpha,M,N}\rangle = 0$ for every polynomial p with degree $[p] \leqq n-1$. First assume that the polynomial p can be written as $p(x) = x^2 \cdot q(x)$ for some polynomial q. Then we have degree $[q] \leqq n-3$ and $n \geqq 3$. In that case we find for $\langle p, L_n^{\alpha,M,N}\rangle$:

$$\begin{aligned}&\frac{1}{\Gamma(\alpha+1)} \cdot \int_0^{+\infty} x^{\alpha+2} e^{-x} \cdot q(x) L_n^{\alpha,M,N}(x)\, dx \\ &= \frac{A_0}{\Gamma(\alpha+1)} \cdot \int_0^{+\infty} x^{\alpha} e^{-x} \cdot x^2 q(x) L_n^{(\alpha)}(x)\, dx \\ &- \frac{A_1}{\Gamma(\alpha+1)} \cdot \int_0^{+\infty} x^{\alpha+1} e^{-x} \cdot x q(x) L_{n-1}^{(\alpha+1)}(x)\, dx \\ &+ \frac{A_2}{\Gamma(\alpha+1)} \cdot \int_0^{+\infty} x^{\alpha+2} e^{-x} \cdot q(x) L_{n-2}^{(\alpha+2)}(x)\, dx,\end{aligned}$$

which equals zero in view of the orthogonality property of the classical Laguerre polynomials.

Now we consider $p(x) = 1$ and $p(x) = x$. First let $p(x) = 1$; then we find for $\langle p, L_n^{\alpha,M,N}\rangle$:

$$\frac{1}{\Gamma(\alpha+1)} \cdot \int_0^{+\infty} x^{\alpha} e^{-x} \cdot L_n^{\alpha,M,N}(x)\, dx + M \cdot L_n^{\alpha,M,N}(0)$$

$$\begin{aligned}&= \frac{A_0}{\Gamma(\alpha+1)} \cdot \int_0^{+\infty} x^{\alpha} e^{-x} \cdot L_n^{(\alpha)}(x)\, dx - \frac{A_1}{\Gamma(\alpha+1)} \cdot \int_0^{+\infty} x^{\alpha} e^{-x} \cdot L_{n-1}^{(\alpha+1)}(x)\, dx \\ &+ \frac{A_2}{\Gamma(\alpha+1)} \cdot \int_0^{+\infty} x^{\alpha} e^{-x} \cdot L_{n-2}^{(\alpha+2)}(x)\, dx \\ &+ M \cdot \left[\binom{n+\alpha}{n} \cdot A_0 - \binom{n+\alpha}{n-1} \cdot A_1 + \binom{n+\alpha}{n-2} \cdot A_2\right].\end{aligned} \tag{3.1}$$

And for $p(x)=x$ we obtain

$$\begin{aligned}
&\frac{1}{\Gamma(\alpha+1)}\cdot\int_0^{+\infty} x^{\alpha+1}\,e^{-x}\cdot L_n^{\alpha,M,N}(x)\,dx+N\cdot\left\{\frac{d}{dx}L_n^{\alpha,M,N}(x)\right\}\bigg|_{x=0}\\
&=\frac{A_0}{\Gamma(\alpha+1)}\cdot\int_0^{+\infty} x^{\alpha+1}\,e^{-x}\cdot L_n^{(\alpha)}(x)\,dx\\
&-\frac{A_1}{\Gamma(\alpha+1)}\cdot\int_0^{+\infty} x^{\alpha+1}\,e^{-x}\cdot L_{n-1}^{(\alpha+1)}(x)\,dx\\
&+\frac{A_2}{\Gamma(\alpha+1)}\cdot\int_0^{+\infty} x^{\alpha+1}\,e^{-x}\cdot L_{n-2}^{(\alpha+2)}(x)\,dx\\
&-N\cdot\left[\binom{n+\alpha}{n-1}\cdot A_0-\binom{n+\alpha}{n-2}\cdot A_1+\binom{n+\alpha}{n-3}\cdot A_2\right].
\end{aligned}\tag{3.2}$$

Using Rodrigues' formula (2.2) and integration by parts we find

$$\int_0^{+\infty} x^{\alpha+k}\,e^{-x}\cdot L_{n-i}^{(\alpha+i)}(x)\,dx=\binom{n-k-1}{n-i}\cdot\Gamma(\alpha+k+1)\tag{3.3}$$

where of course $\binom{p}{q}=0$ for $0\leqq p<q$.

Using this and (1.3) we find for (3.1) and (3.2) by straightforward calculations,

$$-A_1+(n-1)\cdot A_2+M\cdot\left[\binom{n+\alpha}{n}\cdot A_0-\binom{n+\alpha}{n-1}\cdot A_1+\binom{n+\alpha}{n-2}\cdot A_2\right]=0\tag{3.4}$$

and

$$(\alpha+1)\cdot A_2-N\cdot\left[\binom{n+\alpha}{n-1}\cdot A_0-\binom{n+\alpha}{n-2}\cdot A_1+\binom{n+\alpha}{n-3}\cdot A_2\right]=0.\tag{3.5}$$

This proves the orthogonality.

4. Some elementary properties. We start with the coefficients A_0, A_1, and A_2 defined by (1.3). It is not difficult to see that

$$A_0\geqq 1,\quad A_1\geqq 0,\quad\text{and}\quad A_2\geqq 0.$$

For the coefficient k_n of x^n in the polynomial $L_n^{\alpha,M,N}(x)$ we easily find, using (1.2),

$$k_n=\frac{(-1)^n}{n!}\cdot A_0.\tag{4.1}$$

And using (1.2), (2.4), (2.1), (3.4), and (1.3) we obtain

$$M\cdot L_n^{\alpha,M,N}(0)=A_1-(n-1)\cdot A_2=M\cdot\binom{n+\alpha}{n}\cdot\left[1-\frac{N}{(\alpha+1)}\cdot\binom{n+\alpha+1}{n-2}\right].$$

Hence for $M>0$,

$$L_n^{\alpha,M,N}(0)=\binom{n+\alpha}{n}\cdot\left[1-\frac{N}{(\alpha+1)}\cdot\binom{n+\alpha+1}{n-2}\right].\tag{4.2}$$

For $M=0$ we find the same formula by direct calculation.

Note that $L_n^{\alpha,M,N}(0)$ does not depend on M and that $L_n^{\alpha,M,N}(0)\leqq 0$ for $N>0$ and n large enough.

In the same way we obtain, using (1.2), (2.5), (2.1), (3.5), and (1.3),

$$-N\cdot\left\{\frac{d}{dx}L_n^{\alpha,M,N}(x)\right\}\bigg|_{x=0}=(\alpha+1)\cdot A_2=N\cdot\binom{n+\alpha}{n-1}+\frac{MN}{(\alpha+1)}\cdot\binom{n+\alpha}{n}\binom{n+\alpha+1}{n-1}.$$

Hence for $N>0$,

$$\left\{\frac{d}{dx}L_n^{\alpha,M,N}(x)\right\}\bigg|_{x=0}=-\binom{n+\alpha}{n-1}-\frac{M}{(\alpha+1)}\cdot\binom{n+\alpha}{n}\binom{n+\alpha+1}{n-1}. \tag{4.3}$$

For $N=0$ we find the same by direct calculation.

Note that

$$\left\{\frac{d}{dx}L_n^{\alpha,M,N}(x)\right\}\bigg|_{x=0}$$

does not depend on N and that

$$\left\{\frac{d}{dx}L_n^{\alpha,M,N}(x)\right\}\bigg|_{x=0}<0 \quad \text{for } n\geqq 1.$$

Finally, we take $n\geqq 2$. Then we have with (1.1), (1.2), (4.1), (2.4), (3.3), and the orthogonality property,

$$\lambda_n=\langle L_n^{\alpha,M,N}, L_n^{\alpha,M,N}\rangle=k_n\cdot\langle x^n, L_n^{\alpha,M,N}(x)\rangle=\binom{n+\alpha}{n}\cdot A_0\cdot[A_0+A_1+A_2]. \tag{4.4}$$

For $n=0$ and $n=1$ this formula remains valid.

5. Differential equation. In this section we will prove the following.

THEOREM. *The polynomials* $\{L_n^{\alpha,M,N}(x)\}_{n=0}^{+\infty}$ *satisfy a second-order differential equation of the form*

$$x\cdot p_2(x)\cdot\frac{d^2}{dx^2}L_n^{\alpha,M,N}(x)-p_1(x)\cdot\frac{d}{dx}L_n^{\alpha,M,N}(x)+n\cdot p_0(x)\cdot L_n^{\alpha,M,N}(x)=0, \tag{5.1}$$

where $\{p_i(x)\}_{i=0}^{2}$ *are polynomials with*

$$\text{degree}\,[p_0]=\text{degree}\,[p_2]=2 \quad \textit{and} \quad \text{degree}\,[p_1]=3, \tag{5.2}$$

and all having the same leading coefficient which we take to be $(1/n)\cdot A_0\cdot(A_0+A_1+A_2)$.

Proof. We start with the differential equation (2.3) for the classical Laguerre polynomials:

$$x\cdot\frac{d^2}{dx^2}L_n^{(\alpha)}(x)+(\alpha+1-x)\cdot\frac{d}{dx}L_n^{(\alpha)}(x)+n\cdot L_n^{(\alpha)}(x)=0. \tag{5.3}$$

Differentiation of (5.3) gives us

$$x\cdot\frac{d^3}{dx^3}L_n^{(\alpha)}(x)+(\alpha+2-x)\cdot\frac{d^2}{dx^2}L_n^{(\alpha)}(x)+(n-1)\cdot\frac{d}{dx}L_n^{(\alpha)}(x)=0. \tag{5.4}$$

From the definition (1.2) we obtain, using (5.3),

$$L_n^{\alpha,M,N}(x)=q_0(x)\cdot\frac{d}{dx}L_n^{(\alpha)}(x)+q_1(x)\cdot\frac{d^2}{dx^2}L_n^{(\alpha)}(x), \tag{5.5}$$

where

$$\begin{cases} q_0(x)=A_1-\dfrac{1}{n}\cdot A_0\cdot(\alpha+1-x), \\ q_1(x)=A_2-\dfrac{1}{n}\cdot A_0\cdot x. \end{cases} \tag{5.6}$$

Differentiation of (5.5) leads to

$$\begin{aligned} \frac{d}{dx}L_n^{\alpha,M,N}(x)&=q_0'(x)\cdot\frac{d}{dx}L_n^{(\alpha)}(x)+[q_0(x)+q_1'(x)]\cdot\frac{d^2}{dx^2}L_n^{(\alpha)}(x) \\ &\quad+q_1(x)\cdot\frac{d^3}{dx^3}L_n^{(\alpha)}(x). \end{aligned} \tag{5.7}$$

Using (5.4) we find, from (5.7),

$$x\cdot\frac{d}{dx}L_n^{\alpha,M,N}(x)=q_0^*(x)\cdot\frac{d}{dx}L_n^{(\alpha)}(x)+q_1^*(x)\cdot\frac{d^2}{dx^2}L_n^{(\alpha)}(x), \tag{5.8}$$

where

$$\begin{cases} q_0^*(x)=x\cdot q_0'(x)-(n-1)\cdot q_1(x), \\ q_1^*(x)=x\cdot[q_0(x)+q_1'(x)]-(\alpha+2-x)\cdot q_1(x). \end{cases} \tag{5.9}$$

Elimination of the second derivative of the classical Laguerre polynomial in (5.5) and (5.8) leads to

$$\begin{aligned} &[q_0(x)q_1^*(x)-q_0^*(x)q_1(x)]\cdot\frac{d}{dx}L_n^{(\alpha)}(x) \\ &=q_1^*(x)\cdot L_n^{\alpha,M,N}(x)-x\cdot q_1(x)\cdot\frac{d}{dx}L_n^{\alpha,M,N}(x). \end{aligned} \tag{5.10}$$

We define

$$r(x)=q_0(x)q_1^*(x)-q_0^*(x)q_1(x). \tag{5.11}$$

It is easy to check that degree $[r]=2$. Hence $r\not\equiv 0$. So we have from (5.10) and (5.11),

$$\frac{d}{dx}L_n^{(\alpha)}(x)=\frac{q_1^*(x)}{r(x)}\cdot L_n^{\alpha,M,N}(x)-x\cdot\frac{q_1(x)}{r(x)}\cdot\frac{d}{dx}L_n^{\alpha,M,N}(x). \tag{5.12}$$

Substituting (5.12) into (5.5) we obtain a second-order differential equation of the form

$$a_2(x)\cdot\frac{d^2}{dx^2}L_n^{\alpha,M,N}(x)+a_1(x)\cdot\frac{d}{dx}L_n^{\alpha,M,N}(x)+a_0(x)\cdot L_n^{\alpha,M,N}(x)=0, \tag{5.13}$$

where

$$\begin{cases} a_2(x)=x\cdot\{q_1(x)\}^2\cdot r(x), \\ a_1(x)=-q_1(x)\cdot q_1^*(x)\cdot r(x)+x\cdot q_0(x)\cdot q_1(x)\cdot r(x)+\{q_1(x)\}^2\cdot r(x) \\ \qquad +x\cdot q_1(x)\cdot[q_1'(x)\cdot r(x)-q_1(x)\cdot r'(x)], \\ a_0(x)=-q_0(x)\cdot q_1^*(x)\cdot r(x)-q_1(x)\cdot\left[r(x)\cdot\dfrac{d}{dx}q_1^*(x)-q_1^*(x)\cdot r'(x)\right] \\ \qquad +\{r(x)\}^2. \end{cases} \tag{5.14}$$

From (5.11) it follows that

$$q_0(x)\cdot q_1^*(x)-r(x)=q_0^*(x)\cdot q_1(x). \tag{5.15}$$

Hence, the differential equation defined by (5.13) and (5.14) can be divided by $q_1(x)$, which gives us a differential equation of the form (5.13) with

$$\begin{cases} a_2(x)=x\cdot q_1(x)\cdot r(x), \\ a_1(x)=x\cdot q_0(x)\cdot r(x)-q_1^*(x)\cdot r(x)+q_1(x)\cdot r(x) \\ \qquad +x\cdot[q_1'(x)\cdot r(x)-q_1(x)\cdot r'(x)], \\ a_0(x)=q_1^*(x)\cdot r'(x)-r(x)\cdot \dfrac{d}{dx}q_1^*(x)-q_0^*(x)\cdot r(x). \end{cases} \tag{5.16}$$

Using (5.9) we obtain

$$a_1(x)=q_1(x)\cdot[(\alpha+3-x)\cdot r(x)-x\cdot r'(x)]. \tag{5.17}$$

And with (5.6), (5.9), (5.11), and (5.16) we find by straightforward calculations,

$$\begin{aligned} a_0(x)=q_1(x)\cdot\left[q_0^*(x)\cdot\left\{q_0^*(x)+\frac{d}{dx}q_1^*(x)\right\}-q_1^*(x)\cdot\frac{d}{dx}q_0^*(x)\right] \\ +q_1(x)\cdot q_1^*(x)\cdot[(n-1)\cdot\{q_0(x)+q_1'(x)\}-(\alpha+2-x)\cdot q_0'(x)]. \end{aligned}$$

Hence, the differential equation can be divided by $q_1(x)$ once more and we obtain a differential equation of the form (5.13) where

$$\begin{cases} a_2(x)=x\cdot r(x), \\ a_1(x)=(\alpha+3-x)\cdot r(x)-x\cdot r'(x), \\ a_0(x)=q_0^*(x)\cdot\left[q_0^*(x)+\dfrac{d}{dx}q_1^*(x)\right]-q_1^*(x)\cdot\dfrac{d}{dx}q_0^*(x) \\ \qquad +q_1^*(x)\cdot[(n-1)\cdot\{q_0(x)+q_1'(x)\}-(\alpha+2-x)\cdot q_0'(x)]. \end{cases}$$

This proves (5.1) with $p_2(x)=r(x)$, $p_1(x)=-a_1(x)$, and $n\cdot p_0(x)=a_0(x)$.

Now we can easily check that the leading coefficients of p_0, p_1, and p_2 all equal

$$\frac{1}{n}\cdot A_0\cdot(A_0+A_1+A_2).$$

6. Recurrence relation. All classical orthogonal polynomials orthogonal with respect to a weight function satisfy a three term recurrence relation. See, for instance, [3] and [10]. In this section we will prove the following.

THEOREM. *The polynomials* $\{L_n^{\alpha,M,N}(x)\}_{n=0}^{+\infty}$ *satisfy a five term recurrence relation of the form*

$$\begin{aligned} C_{n+2}^{(n)}\cdot L_{n+2}^{\alpha,M,N}(x)+C_{n+1}^{(n)}\cdot L_{n+1}^{\alpha,M,N}(x)+[C_n^{(n)}-x^2]\cdot L_n^{\alpha,M,N}(x) \\ +C_{n-1}^{(n)}\cdot L_{n-1}^{\alpha,M,N}(x)+C_{n-2}^{(n)}\cdot L_{n-2}^{\alpha,M,N}(x)=0. \end{aligned} \tag{6.1}$$

Proof. Since $x^2\cdot L_n^{\alpha,M,N}(x)$ is a polynomial of degree $n+2$ we may write

$$x^2\cdot L_n^{\alpha,M,N}(x)=\sum_{k=0}^{n+2}C_k^{(n)}\cdot L_k^{\alpha,M,N}(x). \tag{6.2}$$

Taking the inner product with $L_i^{\alpha,M,N}(x)$ on both sides of (6.2), we obtain

$$\langle L_i^{\alpha,M,N},L_i^{\alpha,M,N}\rangle\cdot C_i^{(n)}=\langle L_n^{\alpha,M,N}(x),x^2\cdot L_i^{\alpha,M,N}(x)\rangle, \tag{6.3}$$

which follows immediately from the definition of the inner product (1.1). Since $\langle p, L_n^{\alpha,M,N}\rangle = 0$ for every polynomial p with degree $[p] \leqq n-1$, we obtain

$$C_i^{(n)} = 0 \quad \text{for } i = 0, 1, 2, 3, \ldots, n-3. \tag{6.4}$$

This proves the theorem.

From (6.1) it easily follows that

$$C_{n+2}^{(n)} = \frac{k_n}{k_{n+2}} \neq 0, \tag{6.5}$$

where k_n denotes the coefficient of x^n in the polynomial $L_n^{\alpha,M,N}(x)$ given by (4.1). And for $C_{n-2}^{(n)}$ we obtain, from (6.3),

$$C_{n-2}^{(n)} = \frac{k_{n-2} \cdot \lambda_n}{k_n \cdot \lambda_{n-2}} \neq 0, \tag{6.6}$$

where λ_n is defined as in (4.4).

We remark that $C_{n+1}^{(n)}$, $C_n^{(n)}$, and $C_{n-1}^{(n)}$ can be computed by using (6.3), too. We omit the details because these coefficients are not essential here.

7. A Christoffel–Darboux type formula. The classical orthogonal polynomials satisfy the so-called Christoffel-Darboux formula, which gives some information about the zeros of the polynomials. See, for instance, [3] and [10]. In this section we derive an analogous formula for the new polynomials $\{L_n^{\alpha,M,N}(x)\}_{n=0}^{+\infty}$.

From (6.1) we easily obtain

$$\begin{aligned} x^2 \cdot L_{n-i}^{\alpha,M,N}(x) &= C_{n-i+2}^{(n-i)} \cdot L_{n-i+2}^{\alpha,M,N}(x) + C_{n-i+1}^{(n-i)} \cdot L_{n-i+1}^{\alpha,M,N}(x) \\ &\quad + C_{n-i}^{(n-i)} \cdot L_{n-i}^{\alpha,M,N}(x) + C_{n-i-1}^{(n-i)} \cdot L_{n-i-1}^{\alpha,M,N}(x) + C_{n-i-2}^{(n-i)} \cdot L_{n-i-2}^{\alpha,M,N}(x) \end{aligned} \tag{7.1}$$

for $i = 0, 1, 2, 3, \ldots, n-2$.

Now it follows from (1.1) and (7.1) for $0 \leqq i-2 \leqq k \leqq i+2 \leqq n$ that

$$\frac{C_{n-k}^{(n-i)}}{\lambda_{n-i}} = \frac{C_{n-i}^{(n-k)}}{\lambda_{n-k}}. \tag{7.2}$$

From (6.1) and (7.2) we obtain

$$\begin{aligned} &(x^2 - y^2) \cdot \sum_{k=0}^{n} \frac{L_k^{\alpha,M,N}(x) L_k^{\alpha,M,N}(y)}{\lambda_k} \\ &= \frac{C_{n+2}^{(n)}}{\lambda_n} \cdot [L_{n+2}^{\alpha,M,N}(x) L_n^{\alpha,M,N}(y) - L_{n+2}^{\alpha,M,N}(y) L_n^{\alpha,M,N}(x)] \\ &\quad + \frac{C_{n+1}^{(n)}}{\lambda_n} \cdot [L_{n+1}^{\alpha,M,N}(x) L_n^{\alpha,M,N}(y) - L_{n+1}^{\alpha,M,N}(y) L_n^{\alpha,M,N}(x)] \\ &\quad + \frac{C_{n+1}^{(n-1)}}{\lambda_{n-1}} \cdot [L_{n+1}^{\alpha,M,N}(x) L_{n-1}^{\alpha,M,N}(y) - L_{n+1}^{\alpha,M,N}(y) L_{n-1}^{\alpha,M,N}(x)]. \end{aligned} \tag{7.3}$$

Note that we have, with (6.5),

$$C_{n+2}^{(n)} = \frac{k_n}{k_{n+2}} \quad \text{and} \quad C_{n+1}^{(n-1)} = \frac{k_{n-1}}{k_{n+1}}. \tag{7.4}$$

Formula (7.3) can be seen as a Christoffel-Darboux type formula for the polynomials $\{L_n^{\alpha,M,N}(x)\}_{n=0}^{+\infty}$.

Dividing by x^2-y^2 and letting $y\to x$ we obtain the so-called "confluent form":

$$\begin{aligned}
&2x\cdot\sum_{k=0}^{n}\frac{\{L_k^{\alpha,M,N}(x)\}^2}{\lambda_k}\\
&=\frac{k_n}{\lambda_n\cdot k_{n+2}}\cdot\left[L_n^{\alpha,M,N}(x)\cdot\frac{d}{dx}L_{n+2}^{\alpha,M,N}(x)-L_{n+2}^{\alpha,M,N}(x)\cdot\frac{d}{dx}L_n^{\alpha,M,N}(x)\right]\\
(7.5)\qquad&+\frac{C_{n+1}^{(n)}}{\lambda_n}\cdot\left[L_n^{\alpha,M,N}(x)\cdot\frac{d}{dx}L_{n+1}^{\alpha,M,N}(x)-L_{n+1}^{\alpha,M,N}(x)\cdot\frac{d}{dx}L_n^{\alpha,M,N}(x)\right]\\
&+\frac{k_{n-1}}{\lambda_{n-1}\cdot k_{n+1}}\cdot\left[L_{n-1}^{\alpha,M,N}(x)\cdot\frac{d}{dx}L_{n+1}^{\alpha,M,N}(x)-L_{n+1}^{\alpha,M,N}(x)\cdot\frac{d}{dx}L_{n-1}^{\alpha,M,N}(x)\right].
\end{aligned}$$

8. Representation as hypergeometric series. From the definition (1.2) and (2.5) we obtain

$$(8.1)\ L_n^{\alpha,M,N}(x)=\binom{n+\alpha}{n}\cdot\sum_{k=0}^{n}\left[\frac{(-n)_k}{(\alpha+1)_k}\cdot A_0+\frac{(-n)_{k+1}}{(\alpha+1)_{k+1}}\cdot A_1+\frac{(-n)_{k+2}}{(\alpha+1)_{k+2}}\cdot A_2\right]\cdot\frac{x^k}{k!}.$$

If we define

$$(8.2)\qquad L_n^{\alpha,M,N}(x)=\sum_{k=0}^{n}C_k\cdot\frac{x^k}{k!},$$

then it follows from (8.1) and (8.2) that

$$\begin{aligned}
&C_k=\binom{n+\alpha}{n}\cdot\frac{(-n)_k}{(\alpha+1)_{k+2}}\\
(8.3)\qquad&\quad\cdot[A_0\cdot(k+\alpha+1)(k+\alpha+2)+A_1\\
&\quad\cdot(k-n)(k+\alpha+2)+A_2\cdot(k-n)(k-n+1)]\\
&=\binom{n+\alpha}{n}\cdot[A_0+A_1+A_2]\cdot\frac{(-n)_k}{(\alpha+3)_k}\cdot\frac{(k+\beta_0)(k+\beta_1)}{(\alpha+1)_2},
\end{aligned}$$

where

$$(8.4)\quad\begin{cases}[A_0+A_1+A_2]\cdot(\beta_0+\beta_1)=(2\alpha+3)\cdot A_0+(\alpha+2-n)\cdot A_1-(2n-1)\cdot A_2,\\ [A_0+A_1+A_2]\cdot\beta_0\cdot\beta_1=(\alpha+1)(\alpha+2)\cdot A_0-n(\alpha+2)\cdot A_1+n(n-1)\cdot A_2.\end{cases}$$

Since

$$(8.5)\qquad (k+\beta_0)(k+\beta_1)=\beta_0\beta_1\cdot\frac{(\beta_0+1)_k\cdot(\beta_1+1)_k}{(\beta_0)_k\cdot(\beta_1)_k}$$

for $\beta_0\neq 0,-1,-2,-3,\ldots,$ and $\beta_1\neq 0,-1,-2,-3,\ldots,$ we obtain in that case, from (8.2), (8.3), and (8.5),

$$(8.6)\quad L_n^{\alpha,M,N}(x)=\frac{\beta_0\cdot\beta_1}{(\alpha+1)(\alpha+2)}\cdot[A_0+A_1+A_2]\cdot\binom{n+\alpha}{n}\cdot{}_3F_3\left(\begin{matrix}-n,\beta_0+1,\beta_1+1\\ \alpha+3,\beta_0,\beta_1\end{matrix}\middle|\,x\right).$$

With (4.2) we obtain

$$(8.7)[A_0+A_1+A_2]\cdot\frac{\beta_0\cdot\beta_1}{(\alpha+1)(\alpha+2)}=\binom{n+\alpha}{n}^{-1}\cdot L_n^{\alpha,M,N}(0)=1-\frac{N}{(\alpha+1)}\cdot\binom{n+\alpha+1}{n-2}.$$

Hence

$$(8.8)\quad L_n^{\alpha,M,N}(x)=\left[1-\frac{N}{(\alpha+1)}\cdot\binom{n+\alpha+1}{n-2}\right]\cdot\binom{n+\alpha}{n}\cdot{}_3F_3\left(\begin{matrix}-n,\beta_0+1,\beta_1+1\\ \alpha+3,\beta_0,\beta_1\end{matrix}\middle| x\right).$$

Now we examine β_0 and β_1 in somewhat greater detail. First we take $N>0$. The right-hand side of (8.7) is nonpositive for n large enough. So we have with (8.7) and the fact that $A_0+A_1+A_2>0$ for n sufficiently large (for instance),

$$(8.9)\qquad \beta_0\leqq 0 \quad\text{and}\quad \beta_1\geqq 0.$$

Furthermore, we have with (4.3), (8.2), and (8.3),

$$(8.10)\quad \binom{n+\alpha}{n}\cdot[A_0+A_1+A_2]\cdot\frac{(-n)(\beta_0+1)(\beta_1+1)}{(\alpha+1)(\alpha+2)(\alpha+3)}=C_1=\left\{\frac{d}{dx}L_n^{\alpha,M,N}(x)\right\}\Bigg|_{x=0}<0$$

for $n\geqq 1$. Hence for $n\geqq 1$ we have

$$(8.11)\qquad (\beta_0+1)(\beta_1+1)>0.$$

From (8.9) and (8.10) it now follows that for n large enough,

$$(8.12)\qquad -1<\beta_0\leqq 0 \quad\text{and}\quad \beta_1\geqq 0.$$

Since

$$(8.13)\qquad \binom{n+\alpha}{n-i}\sim\frac{n^{\alpha+i}}{\Gamma(\alpha+i+1)}\quad\text{for } n\to+\infty,$$

we find, from (1.3) and (8.13) for $n\to+\infty$, that

$$(8.14)\quad \begin{cases} A_0\sim\dfrac{(\alpha+2)\cdot N}{(\alpha+1)(\alpha+3)}\cdot\dfrac{n^{\alpha+3}}{\Gamma(\alpha+3)} & \text{if } M=0,\\[2ex] A_0\sim\dfrac{M\cdot N}{(\alpha+1)(\alpha+2)}\cdot\dfrac{n^{2\alpha+4}}{\Gamma(\alpha+2)\Gamma(\alpha+4)} & \text{if } M>0,\end{cases}$$

$$(8.15)\quad \begin{cases} A_1\sim\dfrac{N}{(\alpha+1)}\cdot\dfrac{n^{\alpha+2}}{\Gamma(\alpha+2)} & \text{if } M=0,\\[2ex] A_1\sim\dfrac{2MN}{(\alpha+1)^2}\cdot\dfrac{n^{2\alpha+3}}{\Gamma(\alpha+1)\Gamma(\alpha+4)} & \text{if } M>0,\end{cases}$$

$$(8.16)\quad \begin{cases} A_2\sim\dfrac{N}{(\alpha+1)}\cdot\dfrac{n^{\alpha+1}}{\Gamma(\alpha+2)} & \text{if } M=0,\\[2ex] A_2\sim\dfrac{M\cdot N}{(\alpha+1)^2}\cdot\dfrac{n^{2\alpha+2}}{\Gamma(\alpha+1)\Gamma(\alpha+3)} & \text{if } M>0.\end{cases}$$

Hence for $n \to +\infty$ we have

$$(8.17)\quad \begin{cases} (2\alpha+3)\cdot A_0+(\alpha+2-n)\cdot A_1-(2n-1)\cdot A_2 \\ \quad \sim \dfrac{\alpha\cdot N}{(\alpha+1)(\alpha+3)}\cdot\dfrac{n^{\alpha+3}}{\Gamma(\alpha+2)} \qquad (M=0), \\ (2\alpha+3)\cdot A_0+(\alpha+2-n)\cdot A_1-(2n-1)\cdot A_2 \\ \quad \sim -\dfrac{M\cdot N}{(\alpha+1)^2(\alpha+2)}\cdot\dfrac{n^{2\alpha+4}}{\Gamma(\alpha+1)\Gamma(\alpha+4)} \qquad (M>0). \end{cases}$$

It follows from (8.4), (8.7), (8.14)-(8.17) for $n \to +\infty$, that

$$\begin{cases} \beta_0+\beta_1\sim\alpha & \text{if } M=0, \\ \beta_0+\beta_1\sim -1 & \text{if } M>0, \end{cases}$$

and

$$\begin{cases} \beta_0\cdot\beta_1\sim-(\alpha+1) & \text{if } M=0, \\ \beta_0\cdot\beta_1\sim-\dfrac{(\alpha+1)(\alpha+2)\Gamma(\alpha+3)}{M\cdot n^{\alpha+1}} & \text{if } M>0. \end{cases}$$

Hence for $M=0$ we have, for $n\to+\infty$,

$$\beta_0\to-1 \quad\text{and}\quad \beta_1\to\alpha+1$$

and for $M>0$ we have, for $n\to+\infty$,

$$\beta_0\to-1 \quad\text{and}\quad \beta_1\to 0.$$

In the case $N=0$ (Koornwinder's polynomials) we have

$$L_n^{\alpha,M}(x)=\binom{n+\alpha}{n}\cdot{}_2F_2\left(\begin{matrix}-n,\ \gamma+1\\ \alpha+2,\ \gamma\end{matrix}\,\middle|\, x\right),$$

where

$$\gamma=\frac{\alpha+1}{1+M\cdot\dbinom{n+\alpha+1}{n}}>0.$$

Note that for $n\to+\infty$,

$$\begin{cases} \gamma=\alpha+1 & \text{if } M=0, \\ \gamma\to 0 & \text{if } M>0. \end{cases}$$

9. The zeros. All orthogonal polynomials $\{P_n(x)\}_{n=0}^{+\infty}$ which are orthogonal with respect to a weight function have the nice property that the polynomial $P_n(x)$ has n real simple zeros, which are located in the interior of the interval of orthogonality.

Our polynomials $\{L_n^{\alpha,M,N}(x)\}_{n=0}^{+\infty}$ fail to have this property, but we will prove the following.

THEOREM. *The polynomial $L_n^{\alpha,M,N}(x)$ has n real simple zeros. At least $n-1$ of them lie in $(0, +\infty)$, the interior of the interval of orthogonality.*

In other words, at most one zero of $L_n^{\alpha,M,N}(x)$ lies in $(-\infty, 0]$.

Proof. Suppose that $x_1, x_2, \ldots, x_k$ are the zeros of $L_n^{\alpha,M,N}(x)$ which lie in $(0, +\infty)$ and have odd multiplicity. Define

$$p(x) = C \cdot (x-x_1)(x-x_2) \cdots (x-x_k), \qquad C \in \mathbb{R}$$

such that

$$p(x) \cdot L_n^{\alpha,M,N}(x) \geqq 0 \quad \forall x \geqq 0.$$

Define

$$h(x) = (x+d) \cdot p(x)$$

such that $h'(0) = 0$. This implies

$$d = -\frac{p(0)}{p'(0)} > 0,$$

since $p(0)$ and $p'(0)$ have opposite signs.

Hence

$$\langle h, L_n^{\alpha,M,N} \rangle = \frac{1}{\Gamma(\alpha+1)} \cdot \int_0^{+\infty} x^\alpha e^{-x} \cdot h(x) L_n^{\alpha,M,N}(x)\, dx + M \cdot h(0) \cdot L_n^{\alpha,M,N}(0) > 0.$$

Hence degree $[h] \geqq n$, which implies $k \geqq n-1$. This proves the theorem.

Now we examine the nonpositive zero of $L_n^{\alpha,M,N}(x)$ in somewhat greater detail. In view of (4.1) we have $L_n^{\alpha,M,N}(x) > 0$ for $x < 0$ and $|x|$ large enough. This implies that the polynomial $L_n^{\alpha,M,N}(x)$ has a zero in $(-\infty, 0]$ if and only if $L_n^{\alpha,M,N}(0) \leqq 0$. From (4.2) it follows that $L_n^{\alpha,M,N}(0) \leqq 0$ if $N > 0$ and n is sufficiently large. Now we will prove the following.

THEOREM. *Let $N > 0$ and n be sufficiently large such that $L_n^{\alpha,M,N}(x)$ has a zero x_n in $(-\infty, 0]$. Then we have, for $M > 0$,*

$$-\frac{1}{2} \cdot \sqrt{\frac{N}{M}} \leqq x_n \leqq 0. \tag{9.1}$$

Furthermore, we have

$$x_n \to 0 \quad \textit{for } n \to +\infty. \tag{9.2}$$

Proof. Let $x_1, x_2, \ldots, x_{n-1}$ be the zeros of $L_n^{\alpha,M,N}(x)$ which lie in $(0, +\infty)$, and define

$$r(x) = (x-x_1)(x-x_2) \cdots (x-x_{n-1}).$$

Then we may write, in view of (4.1),

$$L_n^{\alpha,M,N}(x) = \frac{(-1)^n}{n!} \cdot A_0 \cdot r(x) \cdot (x-x_n),$$

where $x_n \leqq 0$. Since degree $[r] = n-1$ we have

$$\begin{aligned} 0 = \langle r, L_n^{\alpha,M,N} \rangle &= \frac{(-1)^n \cdot A_0}{n! \cdot \Gamma(\alpha+1)} \cdot \int_0^{+\infty} x^\alpha e^{-x} \cdot r^2(x) \cdot (x-x_n)\, dx \\ &\quad - \frac{(-1)^n}{n!} \cdot A_0 \cdot M \cdot r^2(0) \cdot x_n + \frac{(-1)^n}{n!} \cdot A_0 \cdot N \cdot r'(0) \\ &\quad \cdot [r(0) - x_n \cdot r'(0)]. \end{aligned} \tag{9.3}$$

Since the integral in (9.3) is nonnegative we must have

$$-M\cdot r^2(0)\cdot x_n+N\cdot r'(0)\cdot[r(0)-x_n\cdot r'(0)]\leqq 0.$$

Hence

$$[M\cdot r^2(0)+N\cdot\{r'(0)\}^2]\cdot x_n\geqq N\cdot r(0)\cdot r'(0)=-N\cdot|r(0)\cdot r'(0)|,$$

since $r(0)$ and $r'(0)$ have opposite signs. Now it follows that

$$2\sqrt{M\cdot N}\cdot|r(0)\cdot r'(0)|\cdot x_n\geqq[M\cdot r^2(0)+N\cdot\{r'(0)\}^2]\cdot x_n\geqq -N\cdot|r(0)\cdot r'(0)|.$$

Hence

$$2\sqrt{M\cdot N}\cdot x_n\geqq -N.$$

Now (9.1) follows for $M>0$.

If $x<0$ we may write

$$\begin{aligned}L_n^{\alpha,M,N}(x)&=L_n^{\alpha,M,N}(0)+x\left\{\frac{d}{dx}L_n^{\alpha,M,N}(x)\right\}\bigg|_{x=0}\\&\quad+\frac{x^2}{2!}\cdot\left\{\frac{d^2}{dx^2}L_n^{\alpha,M,N}(x)\right\}\bigg|_{x=0}+\frac{x^3}{3!}\cdot\left\{\frac{d^3}{dx^3}L_n^{\alpha,M,N}(x)\right\}\bigg|_{x=\xi}\end{aligned}\tag{9.4}$$

for some ξ with $x<\xi<0$.

Since the zeros of $(d/dx)L_n^{\alpha,M,N}(x)$ all lie between two consecutive zeros of $L_n^{\alpha,M,N}(x)$ by Rolle's theorem and

$$\left\{\frac{d}{dx}L_n^{\alpha,M,N}(x)\right\}\bigg|_{x=0}<0$$

we must have that all zeros of $(d/dx)L_n^{\alpha,M,N}(x)$ are positive. This means that $(d/dx)L_n^{\alpha,M,N}(x)$ is negative and increasing for $x<0$. In the same way we conclude that $(d^2/dx^2)L_n^{\alpha,M,N}(x)$ must be positive and decreasing for $x<0$. Hence

$$\left\{\frac{d^3}{dx^3}L_n^{\alpha,M,N}(x)\right\}\bigg|_{x=\xi}<0\quad\text{for }\xi<0.$$

Hence, with (9.4),

$$L_n^{\alpha,M,N}(x)>a\cdot x^2+b\cdot x+c\tag{9.5}$$

for $x<0$, where, with (4.2) and (4.3),

$$a=\frac{1}{2}\cdot\left\{\frac{d^2}{dx^2}L_n^{\alpha,M,N}(x)\right\}\bigg|_{x=0},\tag{9.6}$$

$$b=\left\{\frac{d}{dx}L_n^{\alpha,M,N}(x)\right\}\bigg|_{x=0}=-\binom{n+\alpha}{n-1}-\frac{M}{(\alpha+1)}\cdot\binom{n+\alpha}{n}\binom{n+\alpha+1}{n-1}\tag{9.7}$$

and

$$c=L_n^{\alpha,M,N}(0)=\binom{n+\alpha}{n}\cdot\left[1-\frac{N}{(\alpha+1)}\cdot\binom{n+\alpha+1}{n-2}\right].\tag{9.8}$$

Using (1.2), (1.3), (2.1), and (2.4) we obtain

$$\begin{aligned}\left.\left\{\frac{d^2}{dx^2}L_n^{\alpha,M,N}(x)\right\}\right|_{x=0}&=\binom{n+\alpha}{n-2}+\frac{2M}{(\alpha+1)}\cdot\binom{n+\alpha}{n}\binom{n+\alpha+1}{n-2}\\&+\frac{n(\alpha+2)-\alpha}{(\alpha+1)(\alpha+2)(\alpha+4)}\cdot N\cdot\binom{n+\alpha}{n-1}\binom{n+\alpha+1}{n-2}\\&+\frac{2MN}{(\alpha+1)^2(\alpha+2)(\alpha+3)}\cdot\binom{n+\alpha}{n}\binom{n+\alpha+1}{n-1}\binom{n+\alpha+2}{n-2},\end{aligned}\tag{9.9}$$

which can be derived easier by using another representation given by (10.1) and (10.2). With (8.13) it follows from (9.6), (9.7), (9.8), and (9.9) that

$$\begin{cases}a\sim\text{const}\cdot n^{2\alpha+5} & \text{if } M=0,\\ a\sim\text{const}\cdot n^{3\alpha+6} & \text{if } M>0,\end{cases}\tag{9.10}$$

$$\begin{cases}b\sim-\text{const}\cdot n^{\alpha+1} & \text{if } M=0,\\ b\sim-\text{const}\cdot n^{2\alpha+2} & \text{if } M>0,\end{cases}\tag{9.11}$$

and

$$c\sim-\text{const}\cdot n^{2\alpha+3}\tag{9.12}$$

for $n\to+\infty$. The constants in (9.10), (9.11), and (9.12) are all positive. This implies for the sum of the roots of $a\cdot x^2+b\cdot x+c$ that

$$-\frac{b}{2a}\sim\text{const}\cdot n^{-\alpha-4}\to 0\quad\text{if } n\to+\infty,$$

and for the product of the roots,

$$\begin{cases}\dfrac{c}{a}\sim-\text{const}\cdot n^{-2}\to 0 & \text{if } n\to+\infty \text{ in the case } M=0,\\[2ex] \dfrac{c}{a}\sim-\text{const}\cdot n^{-\alpha-3}\to 0 & \text{if } n\to+\infty \text{ in the case } M>0.\end{cases}$$

In view of (9.5), the nonpositive zero x_n lies between the two roots of the "parabola" $a\cdot x^2+b\cdot x+c$. This proves (9.2).

10. Remark. The polynomials $\{L_n^{\alpha,M,N}(x)\}_{n=0}^{+\infty}$ can also be defined by

$$L_n^{\alpha,M,N}(x)=B_0\cdot L_n^{(\alpha)}(x)+B_1\cdot x\cdot L_{n-1}^{(\alpha+2)}(x)+B_2\cdot x^2\cdot L_{n-2}^{(\alpha+4)}(x),\tag{10.1}$$

where

$$\begin{cases}B_0=1-\dfrac{N}{(\alpha+1)}\cdot\dbinom{n+\alpha+1}{n-2},\\[2ex] B_1=-\dfrac{M}{(\alpha+1)}\cdot\dbinom{n+\alpha}{n}-\dfrac{(\alpha+2)\cdot N}{(\alpha+1)(\alpha+3)}\cdot\dbinom{n+\alpha}{n-2},\\[2ex] B_2=\dfrac{N}{(\alpha+1)(\alpha+2)(\alpha+3)}\cdot\dbinom{n+\alpha}{n-1}+\dfrac{M\cdot N}{(\alpha+1)^2(\alpha+2)(\alpha+3)}\\[2ex] \qquad\cdot\dbinom{n+\alpha}{n}\dbinom{n+\alpha+1}{n-1}.\end{cases}\tag{10.2}$$

This can be proved by using the formula

$$n \cdot L_n^{(\alpha)}(x) + (\alpha+1) \cdot \frac{d}{dx} L_n^{(\alpha)}(x) = -x \cdot L_{n-1}^{(\alpha+2)}(x),$$

which follows from the Laguerre differential equation and formula (5.1.13) in [8]. Note that we have

$$\binom{n+\alpha}{n} \cdot B_0 = L_n^{\alpha,M,N}(0),$$

where $B_0 \leqq 0$ if $N > 0$ and n is large enough,

$$A_2 = (\alpha+2)(\alpha+3) \cdot B_2,$$

and

$$B_1 \leqq 0 \quad \text{and} \quad B_2 \geqq 0.$$

Finally, for Koornwinder's polynomials $\{L_n^{\alpha,M}(x)\}_{n=0}^{+\infty}$ this representation yields

$$L_n^{\alpha,M,}(x) = L_n^{(\alpha)}(x) - \frac{M}{(\alpha+1)} \cdot \binom{n+\alpha}{n} \cdot x \cdot L_{n-1}^{(\alpha+2)}(x).$$

REFERENCES

[1] H. Bavinck and H. G. Meijer, *Orthogonal polynomials with respect to a symmetric inner product involving derivatives*, Appl. Anal., 33 (1989), pp. 103-117.

[2] ———, *On orthogonal polynomials with respect to an inner product involving derivatives: zeros and recurrence relations*. Indag. Math., New Series, 1 (1990), pp. 7-14.

[3] T. S. Chihara, *An introduction to orthogonal polynomials*, in Mathematics and Its Applications, Vol. 13, Gordon and Breach, New York, 1978.

[4] R. Koekoek, *Koornwinder's Laguerre polynomials*, Delft Progress Report 12, 1988, pp. 393-404.

[5] ———, *Generalizations of Laguerre polynomials*, J. Math. Anal. Appl., 153 (1990), pp. 576-590.

[6] T. H. Koornwinder, *Orthogonal polynomials with weight function* $(1-x)^\alpha(1+x)^\beta + M \cdot \delta(x+1) + N \cdot \delta(x-1)$, Canad. Math. Bull., 27 (1984), pp. 205-241.

[7] A. M. Krall, *Orthogonal polynomials satisfying fourth order differential equations*, Proc. Roy. Soc. Edinburgh, 87 (1981), pp. 271-288.

[8] H. L. Krall, *Certain differential equations for Tchebycheff polynomials*, Duke Math. J., 4 (1938), pp. 705-718.

[9] ———, *On orthogonal polynomials satisfying a certain fourth order differential equation*, The Pennsylvania State College Studies, University Park, PA, 6, 1940.

[10] G. Szegö, *Orthogonal Polynomials*, American Mathematical Society, Colloquium Publications, 23, 4th ed., Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 24, No. 3, pp. 783–794, May 1993

015

ON MULTIVARIATE ORTHOGONAL POLYNOMIALS*

YUAN XU†

Abstract. Orthogonal polynomials in several variables are studied. The results include a new formulation of the recurrence relation, characteri zation of orthogonality of polynomial sequences, an analogy of Christoffel–Darboux formula, and properties of reproducing kernel function.

Key words. multivariate orthogonal polynomials, three-term relation

AMS subject classifications. 42C05, 33C50

1. Introduction.

One of the most important characteristics of orthogonal polynomials in one variable is the three-term recurrence relation. Let $\{p_n(x)\}$ be a sequence of orthonormal polynomials, $p_n(x) = \gamma_n x^n + \cdots$. Then

$$xp_n(x) = a_n p_{n+1}(x) + b_n p_n(x) + c_n p_{n-1}(x), \tag{1.1}$$

where $p_{-1}(x) \equiv 0$, $a_n = \gamma_n/\gamma_{n+1}$ and $c_n = a_{n-1}$. One application of this relation leads to the Christoffel-Darboux formula

$$\sum_{k=0}^{n} p_k(x)p_k(y) = \frac{\gamma_n}{\gamma_{n+1}} \frac{p_{n+1}(x)p_n(y) - p_n(x)p_{n+1}(y)}{x-y}, \tag{1.2}$$

which plays a fundamental role in the study of orthogonal expansions, since

$$K_n(x,y) := \sum_{k=0}^{n} p_k(x)p_k(y) \tag{1.3}$$

represents the kernel of the partial sums of the orthogonal expansions.

The purpose of this paper is to study the corresponding formulas for multivariate orthogonal polynomials. There are only a few papers dealing with general theory of multivariate orthogonal polynomials (cf. [4], [8]–[11]). One of the difficulties lies in the following fact. For $n > 0$, let V_n be the set of polynomials of total degree n that are orthogonal to all polynomials of lower degree together with zero; then V_n is a vector space of dimension greater than one; there is no reason to assume that one basis for V_n is better than all others. There has been speculation that a general theory should be given for the class V_n rather than some basis of V_n (cf. [7], [11]). In this paper we shall try to follow this point of view.

In his important work [8]–[10], Kowalski introduced a matrix-vector notation to study the recurrence relation; he also characterized the orthogonality of a sequence of polynomials in several variables via recurrence relations. We shall state his main results in §2 after introducing the notation. In §3, we shall present another formulation of the recurrence relation; some important properties of the coefficient matrices will come out of the new formulation. These properties are then used to prove a simplified version of Kowalski's theorem. Moreover, an analogy of the Christoffel–Darboux

*Received by the editors February 14, 1991; accepted for publication August 31, 1992.

†Department of Mathematics, University of Texas, Austin, Texas 78712. Present address: Department of Mathematics, University of Oregon, Eugene, Oregon 97403-1222.

formula will be derived from the recurrence relation in §4; this analogy can be seen as given in terms of V_n rather than a particular basis of V_n. In §5, kernel function and some of its properties will be discussed. Finally, some examples will be given in §6. The work along this line will be continued in our future communications.

This work is inspired by a special recurrence relation for orthogonal polynomials in two variables that has been used to study minimal cubature formulas (cf. [2], [12]).

2. Preliminary. Let Π_n^d be the space of polynomials of total degree n in d variables, and let Π^d be the space of all polynomials in d variables. In all our notation, the superscript d will always associate with number of variables; we shall omit d from any notation after its initial definition, when it causes no confusion. A real valued linear functional $\mathcal{L}$ is said to be a quasi-inner product in Π^d if there exists a basis B of Π^d such that

$$\mathcal{L}(PQ)\begin{cases} =0 & \text{if } P\neq Q \\ \neq 0 & \text{if } P=Q \end{cases} \quad \forall\ P,Q\in B. \tag{2.1}$$

Examples include any linear functional expressible as an integral against a nonnegative weight function w, which induces the inner product $\langle\cdot,\cdot\rangle$,

$$\langle f,g\rangle=\mathcal{L}(fg)=\int_{\mathbb{R}^d} f(\mathbf{x})g(\mathbf{x})w(\mathbf{x})\,d\mathbf{x}. \tag{2.2}$$

For an inner product we take $\mathcal{L}(P^2)=1$ in (2.1).

Given a quasi-inner product $\mathcal{L}$ in Π^d, two polynomials P and Q are said to be orthogonal if $\mathcal{L}(PQ)=0$. For each $k\geq 0$, let $V_k^d\subset\Pi_k^d$ be the set of polynomials of total degree k that are orthogonal to all polynomials in Π_{k-1} together with zero. Then V_k^d is a vector space of dimension $r_k^d := \binom{k+d-1}{k}$. Clearly

$$\Pi_n=\bigoplus_{k=0}^{n} V_k \quad\text{and}\quad \Pi=\bigoplus_{k=0}^{\infty} V_k\ , \tag{2.3}$$

and V_k's are mutually orthogonal. We shall say that $\{V_k\}_{k=0}^{\infty}$ are orthogonal without refering to $\mathcal{L}$. A basis B of Π^d is called orthogonal if

$$B=\bigoplus_{k=0}^{\infty} B_k, \tag{2.4}$$

where B_k is a basis of V_k for each $k\geq 0$. If a basis of V_k is given by $B_k=\{P_1^k,P_2^k,\ldots,P_{r_k}^k\}$, we define vector

$$\mathbb{P}_k(\mathbf{x})=\left[P_1^k(\mathbf{x}),P_2^k(\mathbf{x}),\ldots,P_{r_k}^k(\mathbf{x})\right]^T. \tag{2.5}$$

Sometimes we shall say that $\mathbb{P}_k$ is a basis of V_k, and $\{\mathbb{P}_k\}_{k=0}^{\infty}$ is an orthogonal basis of Π^d. We shall also use the notation (2.5) for any polynomial sequence $\{P_j^k\}_{j=1}^{r_k}$, where the supscript k always means that p_j^k is of total degree k. For $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$, we always write $\mathbf{x}=(x_1,x_2,\ldots,x_d)^T$, $\mathbf{y}=(y_1,y_2,\ldots,y_d)^T$.

We now state Kowalski's theorem. For $\mathbf{x}\in\mathbb{R}^d$, we define

$$\mathbf{x}\mathbb{P}_k(\mathbf{x})=\left[x_1\mathbb{P}_k^T|x_1\mathbb{P}_k^T|\cdots|x_d\mathbb{P}_k^T\right]^T.$$

Let $M_1, M_2, \ldots, M_d$ be any matrices with identical dimensions; we denote

$$bp\Big(\big[M_1|M_2|\cdots|M_d\big]\Big) = \big[M_1^T|M_2^T|\cdots|M_d^T\big]^T.$$

THEOREM K. *Let $B = \{P_i^k\}_{k=0}^{\infty}{}_{i=1}^{r_k}$ be an arbitrary sequence in Π^d. Then the following statements are equivalent:*

(1) *B is a basis of Π^d and there exists a linear functional $\mathcal{L}$ which defines a quasi-inner product in Π^d and satisfies*

$$\mathcal{L}(1) = 1, \quad \mathcal{L}(\mathbb{P}_k\mathbb{P}_\ell^T) = 0, \quad k \neq \ell.$$

(2) *For each $k \geq 0$, there exist (unique) matrices A_k, B_k, and C_k such that*

(a) $\operatorname{rank} A_k = r_{k+1}$;

(b) $\mathbf{x}\mathbb{P}_k = A_k\mathbb{P}_{k+1} + B_k\mathbb{P}_k + C_k\mathbb{P}_{k-1}$ $(C_0 = \mathbb{O}\ ,\ \mathbb{P}_{-1} = \mathbb{O})$;

and for an arbitrary sequence $D_0, D_1, \ldots,$ of matrices satisfying $D_kA_k = I$ (unit matrix) the recursion

(c) $I_0 = [1]$, $I_{k+1} = D_k bp(I_kC_{k+1}^T)$, $k = 0, 1, \ldots,$ *produces nonsingular symmetric matrices I_k.*

The notation $A : i \times j$ means that A is an $i \times j$ matrix. We note that the matrices in the recurrence relation (b) are of the dimensions $A_k : dr_k \times r_{k+1}$, $B_k : dr_k \times r_k$ and $C_k : dr_k \times r_{k+1}$.

3. Recurrence relation. For a matrix $P = (p_{ij})$ whose coefficients p_{ij} belong to Π^d, we denote by $\mathcal{L}(P)$ the matrix whose coefficients are numbers $\mathcal{L}(p_{ij})$. If $\{\mathbb{P}_k(\mathbf{x})\}_{k=0}^{\infty}$ is an orthogonal basis, we denote by H_k the matrix

$$H_k = \mathcal{L}(\mathbb{P}_k\mathbb{P}_k^T). \tag{3.1}$$

From the definition of the quasi-inner product, it follows that H_k is invertible. Indeed, the linear property of $\mathcal{L}$ and (2.1) imply that $\mathcal{L}(P^2) \neq 0$, $\forall\ P \neq 0$, $P \in \Pi^d$. Thus for any vector $\mathbf{c} \in \mathbb{R}^k$,

$$\mathbf{c}^T H_k \mathbf{c} = \mathcal{L}\big([\mathbf{c}^T\mathbb{P}_k]^2\big) \neq 0,$$

which implies that H_k is invertible.

THEOREM 1. *If $\{\mathbb{P}_k(\mathbf{x})\}_{k=0}^{\infty}$ is an orthogonal basis with respect to a quasi-inner product $\mathcal{L}$, $\mathbf{x} \in \mathbb{R}^d$, then for each $k \geq 0$, $1 \leq i \leq d$, there exist unique matrices $A_{k,i} : r_k \times r_{k+1}$, $B_{k,i} : r_k \times r_k$, $C_{k,i} : r_k \times r_{k-1}$ such that*

$$x_i\mathbb{P}_k = A_{k,i}\mathbb{P}_{k+1} + B_{k,i}\mathbb{P}_k + C_{k,i}\mathbb{P}_{k-1}, \qquad 1 \leq i \leq d\ ,\ k \geq 0, \tag{3.2}$$

where $\mathbb{P}_{-1} = 0$ and for all $1 \leq i \leq d$, $k \geq 0$,

$$A_{k,i}H_{k+1} = \mathcal{L}(x_i\mathbb{P}_k\mathbb{P}_{k+1}^T), \tag{3.3}$$

$$B_{k,i}H_k = \mathcal{L}(x_i\mathbb{P}_k\mathbb{P}_k^T), \tag{3.4}$$

$$A_{k,i}H_{k+1} = H_kC_{k+1,i}^T. \tag{3.5}$$

Furthermore, there exist matrices $D_{k,i} : r_{k+1} \times r_k$, $1 \leq i \leq d$, $E_k : r_{k+1} \times r_k$, and $F_k : r_{k+1} \times r_{k-1}$ such that

$$\mathbb{P}_{k+1} = \sum_{i=1}^{d} x_iD_{k,i}\mathbb{P}_k + E_k\mathbb{P}_k + F_k\mathbb{P}_{k-1}, \tag{3.6}$$

where

$$\sum_{i=1}^{d} D_{k,i}A_{k,i} = I, \tag{3.7}$$

$$\sum_{i=1}^{d} D_{k,i}B_{k,i} = -E_k, \qquad \sum_{i=1}^{d} D_{k,i}C_{k,i} = -F_k. \tag{3.8}$$

Proof. The existence of $A_{k,i}$, $B_{k,i}$, and $C_{k,i}$ follows easily from the orthogonality. Actually, (3.2) is just another formulation of Theorem K(b). Multiplying (3.2) by $\mathbb{P}_{k+1}^T$, $\mathbb{P}_k^T$, and $\mathbb{P}_{k-1}^T$ respectively and applying the quasi-inner product, we can easily check that (3.3), (3.4), and (3.5) hold. Equation (3.6) is a reformulation of the corresponding recursion formula of Kowalski [8, Thm. 2]. Substituting $x_i\mathbb{P}_k$ in (3.2) into (3.6), we get (3.7) and (3.8) from orthogonality. □

In the following, we denote by $A_k : dr_k \times r_{k+1}$ and $C_k : r_k \times dr_{k-1}$

$$A_k = \left[A_{k,1}^T | A_{k,2}^T | \cdots | A_{k,d}^T\right]^T \quad \text{and} \quad C_k = \left[C_{k,1} | C_{k,2} | \cdots | C_{k,d}\right]. \tag{3.9}$$

COROLLARY 1. *Let $A_{k,i}$, $C_{k,i}$, $1 \le i \le d$, $k \ge 0$, be as in Theorem 1. Then*

$$\operatorname{rank} A_k = r_{k+1}, \qquad \operatorname{rank} C_{k+1} = r_{k+1} \tag{3.10}$$

and

$$\operatorname{rank} A_{k,i} = \operatorname{rank} C_{k+1,i} = r_k\ , \qquad 1 \le i \le d. \tag{3.11}$$

Proof. Comparing (3.2) and Theorem K(b), we see that A_k in (3.9) is the same as A_k in Theorem K. Thus $\operatorname{rank} A_k = r_{k+1}$. From (3.5) we get that

$$A_k H_{k+1} = G_k C_{k+1}^T,$$

where $G_k : dr_k \times dr_k$ has H_k as diagonal blocks, $G_k = \operatorname{diag}(H_k, H_k, \ldots, H_k)$. Since H_k is invertible, we then have that G_k is invertible; thus

$$\operatorname{rank} C_{k+1} = \operatorname{rank} G_k C_{k+1}^T = \operatorname{rank} A_k H_{k+1} = \operatorname{rank} A_k,$$

which proves (3.10). We now prove (3.11). By (3.3) and (3.5), we only need to prove that $\mathcal{L}(x_i\mathbb{P}_k\mathbb{P}_{k+1}^T)$ has rank r_k. Suppose its rank is less than r_k. Then there exists a nonzero vector $\mathbf{c}_k$ such that

$$\mathbf{c}_k^T \mathcal{L}(x_i\mathbb{P}_k\mathbb{P}_{k+1}^T) = 0.$$

This means that $x_iQ \perp V_{k+1}$, where $Q = \mathbf{c}_k^T\mathbb{P}_k \in V_k$. However, $x_iQ \in \Pi_{k+1} = V_{k+1} \oplus \Pi_k$, thus $x_iQ \in \Pi_k$, which is a contradiction to $Q \in V_k$. □

We remark that the matrices $D_{k,i}$ in the recursion formula (3.6) are not unique when $d > 1$. A way of choosing $D_k = [D_{k,1}|D_{k,2}|\cdots|D_{k,d}]$ is given in [8]. It was also shown there that $\operatorname{rank} D_k = r_{k+1}$. However, it seems unlikely that anything can be said about the rank of $D_{k,i}$ (see Example 2 in §6). From (3.7) and the fact that $\operatorname{rank} A_k = r_k$, some $D_{k,i}$ may even be zero for d large. The recursion formula (3.6) may deserve further attention.

We note that the matrices $B_{k,i}H_k$ are symmetric by (3.4). This fact and the equation (3.5) that came out of the present form of the recurrence relation will enable us to derive an analogy of Christoffel–Darboux formula in §4. The importance of (3.10) can be seen in our next theorem, which simplifies Theorem K.

THEOREM 2. *Let* $B = \{P_j^k\}_{k=0}^{\infty}{}_{j=1}^{r_k}$, $P_1^0 \neq 0$, *be an arbitrary sequence in* Π^d. *Then the following statements are equivalent*:

(1) *There exists a linear function* $\mathcal{L}$ *which defines a quasi-inner product in* Π^d *and makes* $\{\mathbb{P}_k\}_{k=0}^{\infty}$ *an orthogonal basis in* Π^d;

(2) *For* $k \geq 0$, $1 \leq i \leq d$, *there exist matrices* $A_{k,i}$, $B_{k,i}$ *and* $C_{k,i}$ *such that*

(a) $\operatorname{rank} A_k = r_{k+1}$, $\operatorname{rank} C_{k+1} = r_{k+1}$;

(b) $x_i\mathbb{P}_k = A_{k,i}\mathbb{P}_{k+1} + B_{k,i}\mathbb{P}_k + C_{k,i}\mathbb{P}_{k-1}, \quad 1 \leq i \leq d.$

Proof. Clearly the fact that (1) implies (2) is contained in Theorem 1 and its corollary. To prove that (2) implies (1), we shall follow the outline of the proof in [10]. The linear functional

$$\mathcal{L}(1) = 1, \quad \mathcal{L}(\mathbb{P}_k) = 0, \quad k \geq 1$$

is well-defined on Π^d, since $\operatorname{rank} A_k = r_{k+1}$ implies that B is a basis of Π^d. We now use induction to prove that

$$\mathcal{L}(\mathbb{P}_i\mathbb{P}_j^T) = 0, \qquad i \neq j. \tag{3.12}$$

Let $k \geq 0$ be an integer. Assume that (3.12) holds for every i, j such that $0 \leq i \leq k$ and $j > i$. Note that $\operatorname{rank} A_k = r_{k+1}$ and (b) implies that there exist $D_{k,i}$ such that (3.6) holds. Therefore, for $\ell > k+1$,

$$\begin{aligned}\mathcal{L}(\mathbb{P}_{k+1}\mathbb{P}_\ell^T) =& \mathcal{L}\left(\sum_{i=1}^d x_i D_{k,i}\mathbb{P}_k\mathbb{P}_\ell^T\right)\\ =& \mathcal{L}\left(\sum_{i=1}^d D_{k,i}\mathbb{P}_k(A_{\ell,i}\mathbb{P}_{\ell+1} + B_{\ell,i}\mathbb{P}_\ell + C_{\ell,i}\mathbb{P}_{\ell-1})^T\right) = 0.\end{aligned}$$

The induction is complete. Next we prove that $H_k := \mathcal{L}(\mathbb{P}_k\mathbb{P}_{k+1}^T)$ is invertible. Clearly H_k is symmetric. From (b) and (3.12), we get that

$$A_{k,i}H_{k+1} = H_kC_{k+1,i}^T,$$

thus

$$A_kH_{k+1} = G_kC_{k+1}^T,$$

where $G_k = \operatorname{diag}(H_k, \ldots, H_k)$. Since $\mathcal{L}(1) = 1$, we see that $H_0 = \mathcal{L}(\mathbb{P}_0\mathbb{P}_0^T) = (P_1^0)^2$. Thus H_0 is invertible. Suppose H_k is invertible. Then G_k is invertible and by $\operatorname{rank} C_{k+1} = r_{k+1}$ we get

$$\operatorname{rank} A_kH_{k+1} = \operatorname{rank} G_kC_{k+1}^T = r_{k+1}.$$

However, $\operatorname{rank} A_k = r_{k+1}$ by (a) and $A_k : dr_k \times r_{k+1}$, $H_{k+1} : r_{k+1} \times r_{k+1}$; we then have (see [6, p. 66])

$$\begin{aligned}\operatorname{rank} H_{k+1} \geq \operatorname{rank}(A_kH_{k+1}^T) &\geq \operatorname{rank} A_k + \operatorname{rank} H_{k+1}^T - r_{k+1}\\ &= \operatorname{rank} H_{k+1}^T = \operatorname{rank} H_{k+1}.\end{aligned}$$

Therefore,

$$\operatorname{rank} H_{k+1} = \operatorname{rank} A_kH_{k+1}^T = r_{k+1},$$

which implies that H_{k+1} is invertible. By induction, we have then proved that H_k is invertible for each $k \geq 0$. Since H_k is symmetric and invertible there exist invertible matrices S_k and I_k such that $H_k = S_k I_k S_k^T$ and I_k is diagonal. For $\mathbb{Q}_k = S_k^{-1}\mathbb{P}_k$ we then have

$$\mathcal{L}(\mathbb{Q}_k\mathbb{Q}_k^T) = S_k^{-1}\mathcal{L}(\mathbb{P}_k\mathbb{P}_k^T)(S_k^{-1})^T = S_k^{-1}H_k(S_k^{-1})^T = I_k.$$

This proves that $\mathcal{L}$ defines a quasi-inner product in Π^d; $\mathcal{L}$ makes $\{\mathbb{P}_k\}_{k=0}^{\infty}$ an orthogonal basis by (3.12). The proof is completed. □

This theorem should be compared with that of Favard for polynomials in one variable (cf. [5, p.18–22]). We note that a comparison criterion of orthogonality for sequences of polynomials and an integral representation for a corresponding quasi-inner product are given in [9].

4. Christoffel–Darboux formula. The following theorem is a generalization of the well-known Christoffel–Darboux formula (1.3) (cf. [5, p.23]).

THEOREM 3. *Let* $\mathbf{x} \in \mathbb{R}^d$ *and* $\{\mathbb{P}_k(\mathbf{x})\}_{k=0}^{\infty}$ *satisfy*

$$(4.1) \qquad x_i\mathbb{P}_k = A_{k,i}\mathbb{P}_{k+1} + B_{k,i}\mathbb{P}_k + C_{k,i}\mathbb{P}_{k-1}, \qquad 1 \leq i \leq d,$$

where $\mathbb{P}_{-1} = 0$ *and*

$$(4.2) \qquad A_{k,i}H_{k+1} = H_kC_{k+1,i}^T, \qquad 1 \leq i \leq d,\ k \geq 0,$$

with H_k *being symmetric and invertible and* $B_{k,i}H_k$ *being symmetric. Then for any integer* $n \geq 0$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$(4.3) \sum_{k=0}^{n} \mathbb{P}_k^T(\mathbf{x})H_k^{-1}\mathbb{P}_k(\mathbf{y}) = \frac{\left[A_{n,i}\mathbb{P}_{n+1}(\mathbf{x})\right]^T H_n^{-1}\mathbb{P}_n(\mathbf{y}) - \mathbb{P}_n^T(\mathbf{x})H_n^{-1}\left[A_{n,i}\mathbb{P}_{n+1}(\mathbf{y})\right]}{x_i - y_i},$$

$$1 \leq i \leq d.$$

Proof. Let

$$\Sigma_k = \left[A_{k,i}\mathbb{P}_{k+1}(\mathbf{x})\right]^T H_k^{-1}\mathbb{P}_k(\mathbf{y}) - \mathbb{P}_k(\mathbf{x})^T H_k^{-1}\left[A_{k,i}\mathbb{P}_{k+1}(\mathbf{y})\right].$$

From the recurrence relation (4.1) we have

$$(4.4) \qquad \begin{aligned} \Sigma_k &= \left[x_i\mathbb{P}_k(\mathbf{x}) - B_{k,i}\mathbb{P}_k(\mathbf{x}) - C_{k,i}\mathbb{P}_k(\mathbf{x})\right]^T H_k^{-1}\mathbb{P}_k(\mathbf{y}) \\ &\quad - \mathbb{P}_k^T(\mathbf{x})H_k^{-1}\left[y_i\mathbb{P}_k(\mathbf{y}) - B_{k,i}\mathbb{P}_k(\mathbf{y}) - C_{k,i}\mathbb{P}_{k-1}(\mathbf{y})\right] \\ &= (x_i - y_i)\mathbb{P}_k^T(\mathbf{x})H_k^{-1}\mathbb{P}_k(\mathbf{y}) - \mathbb{P}_k^T(\mathbf{x})\left[B_{k,i}^T H_k^{-1} - H_k^{-1}B_{k,i}\right]\mathbb{P}_k(\mathbf{y}) \\ &\quad - \left[\mathbb{P}_{k-1}^T(\mathbf{x})C_{k,i}^T H_k^{-1}\mathbb{P}_k(\mathbf{y}) - \mathbb{P}_k^T(\mathbf{x})H_k^{-1}C_{k,i}\mathbb{P}_{k-1}(\mathbf{y})\right]. \end{aligned}$$

Since H_k and $B_{k,i}H_k$ are both symmetric, we have

$$B_{k,i}H_k = (B_{k,i}H_k)^T = H_kB_{k,i}^T,$$

which implies that

$$B_{k,i}^T H_k^{-1} = H_k^{-1}B_{k,i}.$$

Therefore, the second term in the right-hand side of (4.4) is zero. From (4.2) we have that

$$C_{k,i}^T H_k^{-1} = H_{k-1}^{-1}A_{k-1,i}.$$

Thus the third term in the right-hand side of (4.4) is

$$-\mathbb{P}_{k-1}^T(\mathbf{x})H_k^{-1}A_{k-1,i}\mathbb{P}_k(\mathbf{y}) + \mathbb{P}_k^T(\mathbf{x})A_{k-1,i}^T H_{k-1}^{-1}\mathbb{P}_{k-1}(\mathbf{y}) = \Sigma_{k-1}.$$

Therefore, (4.4) can be rewritten as

$$(x_i - y_i)\mathbb{P}_k^T(\mathbf{x})H_k^{-1}\mathbb{P}_k(\mathbf{y}) = \Sigma_k - \Sigma_{k-1}.$$

Summing this identity from zero to n and noticing that $\Sigma_{-1} = 0$, we obtain (4.3). □

COROLLARY. *Let $\{\mathbb{P}_k(\mathbf{x})\}_{k=0}^{\infty}$ be as in Theorem 1. Then*

(4.5)
$$\sum_{k=0}^{n}\mathbb{P}_k^T(\mathbf{x})H_k^{-1}\mathbb{P}_k(\mathbf{x}) = \mathbb{P}_n^T(\mathbf{x})H_n^{-1}\left[A_{n,i}\partial_i\mathbb{P}_{n+1}(\mathbf{x})\right] - \left[A_{n,i}\mathbb{P}_{n+1}(\mathbf{x})\right]^T H_n^{-1}\partial_i\mathbb{P}_n(\mathbf{x}),$$

$$1 \le i \le d,$$

where $\partial_i = \partial/\partial x_i$ denotes the partial derivative with respect to x_i.

Proof. Since $\mathbb{P}_n(\mathbf{x})^T H_n^{-1}[A_{n,i}\mathbb{P}_{n+1}(\mathbf{x})]$ is a scalar function, it is equal to its own transpose. Thus,

$$\mathbb{P}_n(\mathbf{x})^T H_n^{-1}\left[A_{n,i}\mathbb{P}_{n+1}(\mathbf{x})\right] = \left[A_{n,i}\mathbb{P}_{n+1}(\mathbf{x})\right]^T H_n^{-1}\mathbb{P}_n(\mathbf{x}).$$

Therefore, the numerator of the right-hand side of (4.3) can be written as

$$\left[A_{n,i}\mathbb{P}_{n+1}(\mathbf{x})\right]^T H_n^{-1}\left[\mathbb{P}_n(\mathbf{y}) - \mathbb{P}_n(\mathbf{x})\right] - \mathbb{P}_n(\mathbf{x})^T H_n^{-1}A_{n,i}\left[\mathbb{P}_{n+1}(\mathbf{y}) - \mathbb{P}_{n+1}(\mathbf{x})\right].$$

Thus (4.5) follows from (4.3) by letting $y_i \to x_i$. □

In particular, if H_k is positive definite then the right-hand side of (4.5) is positive for all $\mathbf{x} \in \mathbb{R}^d$. The important examples include all orthogonal polynomials with respect to inner product (2.2). When $\{\mathbb{P}_k(\mathbf{x})\}_{k=0}^{\infty}$ in Theorem 3 is an orthogonal basis, $\mathbb{P}_k(\mathbf{x})$ is a basis for vector space V_k for each $k \ge 0$. However, (4.3) is actually independent of this particular basis of V_k. Let

$$K_n(\mathbf{x},\mathbf{y}) = \sum_{k=0}^{n-1}\mathbb{P}_k^T(\mathbf{x})H_k^{-1}\mathbb{P}_k(\mathbf{y}). \tag{4.6}$$

THEOREM 4. *Let V_k, $k \ge 0$, be defined via a quasi-inner product $\mathcal{L}$. Then $K_n(\cdot,\cdot)$ depends only on V_k rather than a particular basis of V_k.*

Proof. Let $\mathbb{P}_k$ be a basis of V_k. If $\mathbb{Q}_k$ is another basis, then there exists an invertible scalar matrix M_k such that

$$\mathbb{P}_k(\mathbf{x}) = M_k\mathbb{Q}_k(\mathbf{x}).$$

Let $H_k = H_k(\mathbb{P}) = \mathcal{L}(\mathbb{P}_k\mathbb{P}_k^T)$ and $H_k(\mathbb{Q}) = \mathcal{L}(\mathbb{Q}_k\mathbb{Q}_k^T)$. Then

$$H_k^{-1}(\mathbb{P}) = M_k^{-1}H_k^{-1}(\mathbb{Q})(M_k^T)^{-1}.$$

Therefore, we have

$$\mathbb{P}_k^T H_k^{-1}(\mathbb{P})\mathbb{P}_k = \mathbb{Q}_k^T H_k^{-1}(\mathbb{Q})\mathbb{Q}_k,$$

which proves the desired result. □

As a consequence of this theorem, the formula (4.3) can be seen as a natural extension of the Christoffel–Darboux formula of one variable, since the orthogonality is defined via V_k (cf. (2.4)). From this theorem, one can choose some convenient basis of V_k when dealing with $K_n(\cdot,\cdot)$. In particular, if an orthonormal basis of V_k is chosen, then there will be no inverse matrix operation in the formula, since H_k becomes an identity matrix. The choice of orthonormal basis is particularly preferable when we need to prove certain properties of $K_n(\cdot,\cdot)$, as we shall see in the next section.

We note that another expression of the Christoffel–Darboux formula is given in [4]. However, the formula is much more complicated; in particular, the numerator of the right-hand side in it has more terms than the left-hand side that is defined as (4.6).

5. Kernel function. For orthogonal polynomials in one variable, $K_n(\cdot,\cdot)$ is the reproducing kernel function; it has many important properties and applications. We now discuss some of its corresponding properties for multivariate orthogonal polynomials.

THEOREM 5. *Let $\mathcal{L}$ be an inner product in Π^d. Then*

(1) *$K_n(\cdot,\cdot)$ is the reproducing kernel, i.e., for all $P \in \Pi_{n-1}$,*

$$P(\mathbf{x}) = \mathcal{L}\big[K_n(\mathbf{x},\cdot)P(\cdot)\big]; \tag{5.1}$$

(2) *For an arbitrary point $\mathbf{t} \in \mathbb{R}^d$,*

$$K_n(\mathbf{t},\mathbf{t})^{-1} = \min \mathcal{L}[P^2], \tag{5.2}$$

where minimum is over all polynomials $P \in \Pi_{n-1}$ subject to the condition $P(\mathbf{t}) = 1$.

Proof. Since $\mathcal{L}$ is an inner product, by (2.1) there exists a basis $B = \{P_j^k\}_{k=0}^{\infty}{}_{j=0}^{r_k}$ of Π^d such that

$$\mathcal{L}(P_i^k P_j^\ell) = \delta_{ij}\delta_{k\ell}. \tag{5.3}$$

For any polynomial $P \in \Pi_{n-1}$, we have by orthogonality

$$P(\mathbf{x}) = \sum_{k=0}^{n-1} \mathbb{P}_k(\mathbf{x})^T \mathbf{a}_k(P), \quad \text{where } \mathbf{a}_k(P) = \mathcal{L}(P\mathbb{P}_k) \tag{5.4}$$

or

$$P(\mathbf{x}) = \sum_{k=0}^{n-1} \mathbb{P}_k(\mathbf{x})^T \mathcal{L}(P\mathbb{P}_k) = \mathcal{L}\big[K_n(\mathbf{x},\cdot)P(\cdot)\big],$$

which is (5.1). From (5.3) and (5.4), we also have

$$\mathcal{L}[P^2] = \sum_{k=0}^{n-1} \mathbf{a}_k(P)^T \mathbf{a}_k(P) = \sum_{k=0}^{n-1} \|\mathbf{a}_k(P)\|^2 ,$$

where $\|\cdot\|$ is the Euclidean norm, $\|\mathbf{a}\|^2 = \mathbf{a}^T\mathbf{a}$. If $P(\mathbf{t}) = 1$, we then have by Cauchy's inequality,

$$1 = P(\mathbf{t}) = \sum_{k=0}^{n-1} \mathbb{P}_k(\mathbf{t})^T \mathbf{a}_k(P) \le \sum_{k=0}^{n-1} \|\mathbf{a}_k(P)\| \, \|\mathbb{P}_k(\mathbf{t})\|$$

$$\le \sum_{k=0}^{n-1} \|\mathbf{a}_k(P)\|^2 \sum_{k=0}^{n-1} \|P_k(\mathbf{t})\|^2 = \mathcal{L}[P^2]K_n(\mathbf{t},\mathbf{t}),$$

where equality holds if and only if

$$\mathbf{a}_k(P) = KP_k(\mathbf{t}), \qquad K = [K_n(\mathbf{t},\mathbf{t})]^{-1}.$$

This proves (5.2), and the proof is completed. □

These two properties are completely analogous to the corresponding ones in one variable (cf. [5, p.38]); their proofs are similar as well. We enclosed the proof here to illustrate the importance of Theorem 4, which allows us to use orthonormal basis in the proof. We now briefly discuss the role of $K_n(\cdot,\cdot)$ as the kernel function of the orthogonal expansion. This part illustrates very well the point of view that we mentioned in the introduction (see also [11]). Most of the results in the following can be obtained from the standard theorems in inner product space.

Now let $\mathcal{L}$ be an inner product expressible in the form of (2.2), or more generally, we replace $w(\mathbf{x})\,d\mathbf{x}$ by some distribution $d\alpha$, $\int d\alpha > 0$. If $f \in L_{d\alpha}$, then by (2.3), the expansion

$$f \sim \sum_{k=0}^{\infty} U_k(f), \qquad U_k(f) = \operatorname{proj}_{V_k} f \tag{5.5}$$

can be formed (at least formally). Let $\mathbb{P}_k$ be a basis of V_k. Then $U_k(f)$ can be expressed by $\mathbb{P}_k$ as

$$U_k(f) = \mathbb{P}_k^T \mathbf{a}_k(f), \qquad \mathbf{a}_k(f) \in \mathbb{R}^{r_k}. \tag{5.6}$$

Therefore, we have

$$\int f\mathbb{P}_k\,d\alpha = \int U_k(f)\mathbb{P}_k\,d\alpha = \int \mathbb{P}_k\mathbb{P}_k^T\mathbf{a}_k(f) = H_k\mathbf{a}_k(f). \tag{5.7}$$

We shall call the expansion

$$f \sim \sum_{k=0}^{\infty} \mathbb{P}_k^T\mathbf{a}_k(f) = \sum_{k=0}^{\infty} \mathbb{P}_k^T H_k^{-1} \int f\mathbb{P}_k\,d\alpha \tag{5.8}$$

the Fourier orthogonal expansion of f. Let $S_nf = S_n(d\alpha, f)$ denote the nth partial sum of (5.8). Then we have from (4.6) that

$$S_nf(\mathbf{x}) = \int_{\mathbb{R}^d} K_n(\mathbf{x},\mathbf{y})f(\mathbf{y})\,d\alpha(\mathbf{y}).$$

Therefore, $K_n(\cdot,\cdot)$ is the kernel function of the partial sums. From Theorem 4, S_nf is defined in terms of V_k, $0 \le k \le n-1$, rather than some basis of V_k. The proof of the following theorem relies on this fact; otherwise, the proof is standard.

THEOREM 6. *Let $\mathcal{L}$ be an inner product associated with $d\alpha$ (or as in (2.2)), $f \in L^2_{d\alpha}$. Then among all polynomials P in Π_{n-1}, the integral*

$$\int |f(\mathbf{x}) - P(\mathbf{x})|^2\,d\alpha$$

becomes minimal if and only if $P = S_n(d\alpha, f)$.

Proof. Let $\mathbb{P}_k$ be an orthonormal basis of V_k. For any $P \in \Pi_{n-1}$ there exists $\mathbf{b}_k$ such that

$$P(\mathbf{x}) = \sum_{k=0}^{n-1} \mathbf{b}_k^T \mathbb{P}_k(\mathbf{x}).$$

Since $S_n f$ is defined in terms of V_k, we have

$$S_n f(\mathbf{x}) = \sum_{k=0}^{n-1} \mathbf{a}_k^T \mathbb{P}_k(\mathbf{x}), \qquad \text{where } \mathbf{a}_k = \int f \mathbb{P}_k \, d\alpha.$$

Then we follow the standard argument to get that

$$0 \le \int |f - P|^2 \, d\alpha = \int f^2 \, d\alpha - 2 \sum \mathbf{b}_k^T \int f \mathbb{P}_k \, d\alpha + \sum_k \sum_j \mathbf{b}_k^T \int \mathbb{P}_k \mathbb{P}_j^T \mathbf{b}_j$$

$$= \int f^2 \, d\alpha - 2 \sum_{k=0}^{n-1} \mathbf{b}_k^T \mathbf{a}_k + \sum_{k=0}^{n-1} \mathbf{b}_k^T \mathbf{b}_k$$

$$= \int f^2 \, d\alpha - \sum_{k=0}^{n-1} \mathbf{a}_k^T \mathbf{a}_k + \sum_{k=0}^{n-1} [\mathbf{a}_k^T \mathbf{a}_k + \mathbf{b}_k^T \mathbf{b}_k - 2\mathbf{b}_k^T \mathbf{a}_k] \; .$$

By Cauchy's inequality the third term in the right-hand side is nonnegative; and the integral is minimal if and only if $\mathbf{b}_k = \mathbf{a}_k$ or $P = S_n f$. □

The minimal itself leads to the Bessel's inequality, which, in terms of a basis $\mathbb{P}_k$ of V_k, is

$$\sum_{k=0}^{\infty} \mathbf{a}_k(f)^T H_k^{-1} \mathbf{a}_k(f) \le \int |f|^2 \, d\alpha,$$

where $\mathbf{a}_k(f)$ is given by (5.7). We note that $\mathbf{a}_k(f)^T H_k^{-1} \mathbf{a}_k(f)$ does not depend on a particular choice of the basis of V_k. When Π^d is dense in $L^2(d\alpha)$, the Parseval's identity that is (5.9) with equality holds and $S_n f$ converges to f in $L^2(d\alpha)$. For $d > 1$, the convergence behavior of $S_n f$ other than that of $L^2(d\alpha)$ has not been studied; one may start with the special bivariable orthogonal polynomials that have compact formulas such as those given in survey paper [7] by Koornwinder.

6. Examples. All examples in this section are given by polynomials in two variables which are orthogonal with respect to some weight functions w (see (2.2)). The first example deals with the simplest case that the orthogonal polynomials are the product of two univariant ones.

Example 1. Let $P_j^k(x_1, x_2) = p_{j-1}(x_1) q_{k-j+1}(x_2)$, $k \ge 0$, $1 \le j \le k+1$, where $p_j(x) = \gamma_j x^j + \cdots$ and $q_j(x) = \delta_j x^j + \cdots$ are the orthogonal polynomials on the symmetric intervals I_p and I_q with respect to even weight functions w_p and w_q, respectively. Then polynomials $\{P_j^k\}$ are orthogonal over the region $I_p \times I_q$ with respect to weight function $w_p(x_1) w_q(x_2)$. Examples of $\{p_j\}$ and $\{q_j\}$ include ultraspherical polynomials and Hermite polynomials. Let $\Gamma_{j+1} = \gamma_j / \gamma_{j+1}$, $\Delta_{j+1} = \delta_j / \delta_{j+1}$. We can easily verify that the recurrence relation (3.2) is

$$x_i \mathbb{P}_k = A_{k,i} \mathbb{P}_{k+1} + A_{k-1,i}^T \mathbb{P}_{k-1}, \qquad k \ge 0, \qquad i = 1, 2,$$

where

$$A_{k,1} = \big[\mathbb{O} \mid \operatorname{diag}(\Gamma_0, \Gamma_1, \dots, \Gamma_k)\big], \qquad A_{k,2} = \big[\operatorname{diag}(\Delta_k, \Delta_{k-1}, \dots, \Delta_0) \mid \mathbb{O}\big].$$

The recursion formula (3.6) now takes the form

$$\mathbb{P}_{k+1} = (x_1 D_{k,1} + x_2 D_{k,2})\mathbb{P}_k + F_k \mathbb{P}_{k-1},$$

where $D_{k,i}$, which are not unique, can be chosen as

$$D_{k,1} = \big[\mathbb{O} \mid \operatorname{diag}(\tfrac{1}{2}\Gamma_0^{-1}, \dots, \tfrac{1}{2}\Gamma_{k-1}^{-1}, \Gamma_k^{-1})\big]^T,$$
$$D_{k,2} = \big[\operatorname{diag}(\Delta_k^{-1}, \tfrac{1}{2}\Delta_{k-1}^{-1}, \dots, \tfrac{1}{2}\Delta_0^{-1}) \mid \mathbb{O}\big]^T,$$

and F_k are given by (3.8) with $C_{k,i} = A_{k-1,i}^T$.

As we pointed out before, the choice of orthonormal basis for V_k seems to be preferable in view of the Christoffel–Darboux formula. However, there are other bases that have been used in applications because of their other relatively simpler structures. An important example is the basis whose elements are monic polynomials (cf. [2], [11], [12]). Our next example illustrates this choice.

Example 2. Let $S = \{(x_1, x_2) \mid x_1 \geq 0, x_2 \geq 0, x_1 + x_2 \leq 1\}$. Let V_k be orthogonal with respect to the weight function $W_{\alpha,\beta,\gamma}$ that has support on S, $W_{\alpha,\beta,\gamma}(x_1, x_2) = x_1^\alpha x_2^\beta (1 - x_1 - x_2)^\gamma$, $(x_1, x_2) \in S$, $\alpha, \beta, \gamma > -1$.

One orthonormal basis of V_k is given by (cf. [7, p. 449])

$$P_j^k(x_1, x_2) = c_{jk}(1 - x_2)^j p_j^{(\beta,\gamma)}\left(\frac{x_1}{1 - x_2}\right) p_{k-j}^{(\alpha,\beta+\gamma+2j+1)}(x_2).$$

Here $p_j^{(a,b)}(x)$ is the classical Jacobi polynomial on $[0, 1]$ with respect to $w(x) = x^a(1-x)^b$, $c_{jk} \neq 0$ are the proper constants. It can be verified that the coefficient matrices in the recurrence relation (3.2) have the form

$$A_{k,1} = [T \mid \mathbf{a}], \qquad A_{k,2} = [D \mid \mathbb{O}],$$

where T is a tridiagonal matrix, D is a diagonal matrix, and $\mathbf{a} = (0, \dots, 0, a)^T$ for some $a \neq 0$.

Another basis that we consider is the monic one,

$$Q_{j+1}^k(x_1, x_2) = x_1^j x_2^{k-j} + q_j^k(x_1, x_2), \quad \text{where } q_j^k \in \Pi_{k-1},$$

$k \geq 0$, $0 \leq j \leq k$. These orthogonal polynomials are first considered in [1]; they have some very nice properties (see [1]–[3]). In [2], all coefficient matrices in the recurrence relation (3.2) have been given explicitly; the matrices $A_{k,i}$ are simply

$$A_{k,1} = (I_k \mid \mathbb{O}), \qquad A_{k,2} = (\mathbb{O} \mid I_k),$$

and $B_{k,i}$ are tridiagonal matrices, and $C_{k,i}$ are of the form $[T_k \mid \mathbf{a}]$, where T_k's are the tridiagonal matrices and $\mathbf{a} = (0, \dots, 0, a)^T$. We refer to [2] for the explicit formulas for $B_{k,i}$ and $C_{k,i}$. The matrices $D_{k,i}$ in (3.6) can be chosen to be very simple in many ways; however, there does not seem to be a choice that is better than all others. One possible choice is

$$D_{k,1} = (I_k \mid \mathbb{O})^T, \qquad D_{k,2} = (0 \mid \mathbf{e})^T,$$

where $\mathbf{e} = (0, \ldots, 0, 1)^T$. Since the monic orthogonal polynomials play a very important role in the study of the minimal cubature formulas (cf. [2], [12]), a more comprehensive study of them seems to be worthwhile.

Acknowledgment. I thank Professors H. Berens and H. Schmid for making me acquainted with their work and other assistance. I also thank Professor T. Koornwinder for bringing [10] to my attention.

REFERENCES

[1] P. Appell and J. K. de Fériet, *Functions Hypergéometriques et Hypersphériques, Polynomes d'Hermite*, Gauthier-Villars et Cie, Paris, 1926.

[2] H. Berens and H. Schmid, *On the number of nodes of odd degree cubature formulae for integrals with Jacobi weights on a simplex*, in Numerical Integration, T. O. Espelid and A. Genz, eds., Kluwer Academic, Norwell, MA, 1992, pp. 37–44.

[3] H. Berens, H. Schmid, and Y. Xu, *Bernstein–Durrmeyer polynomials on simplicies*, J. Approx. Theory, 68 (1992), pp. 246–260.

[4] M. Bertran, *Note on orthogonal polynomials in ν-variables*, SIAM J. Math. Anal., 6 (1975), pp. 250–257.

[5] T. S. Chihara, *An Introduction to Orthogonal Polynomials*, Mathematics and Its Applications, Vol. 13, Gordon and Breach, New York, 1978.

[6] F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1959.

[7] T. Koornwinder, *Two-variable analogues of the classical orthogonal polynomials*, in Theory and Application of Special Functions, R. A. Askey, ed., Academic Press, New York, 1975.

[8] M. A. Kowalski, *The recursion formulas for orthogonal polynomials in n variables*, SIAM J. Math. Anal., 13 (1982), pp. 309–315.

[9] ———, *Orthogonality and recursion formulas for polynomials in n variables*, SIAM J. Math. Anal., 13 (1982), pp. 316–323.

[10] ———, *Algebraic characterization of orthogonality in the space of polynomials*, Lecture Notes in Math. 1171 (1985), pp. 101–110.

[11] H. L. Krall and I. M. Sheffer, *Orthogonal polynomials in two variables*, Ann. Mat. Pura. Appl., 76 (1967), pp. 325–376.

[12] H. Schmid, *Minimal cubature formula and matrix-equation*, manuscript, 1992.

[13] G. Szegö, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ., 23, Fourth Ed., American Mathematical Society, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 24, No. 3, pp. 795–813, May 1993

ASKEY–WILSON POLYNOMIALS AS ZONAL SPHERICAL FUNCTIONS ON THE $SU(2)$ QUANTUM GROUP*

TOM H. KOORNWINDER[†]

Abstract. On the $SU(2)$ quantum group the notion of (zonal) spherical element is generalized by considering left and right invariance in the infinitesimal sense with respect to twisted primitive elements of the $sl(2)$ quantized universal enveloping algebra. The resulting spherical elements belonging to irreducible representations of quantum $SU(2)$ turn out to be expressible as a two-parameter family of Askey–Wilson polynomials. For a related basis change of the representation space a matrix of dual q-Krawtchouk polynomials is obtained. Big and little q-Jacobi polynomials are obtained as limits of Askey–Wilson polynomials.

Key words. quantum groups, $SU(2)$, spherical functions, infinitesimal invariance, Askey–Wilson polynomials, dual q-Krawtchouk polynomials, big q-Jacobi polynomials, little q-Jacobi polynomials

AMS subject classifications. 33D45, 33D80, 17B37

1. Introduction. One of the interesting aspects of the rapidly developing subject of quantum groups is that they seem to provide the natural setting for q-hypergeometric functions and orthogonal polynomials. For the relatively simple case of $SU_q(2)$ a first example of this phenomenon was given by Vaksman and Soĭbel'man [21] (see also Masuda et al. [14], [13] and the author [10]), where it was shown that the matrix elements of the irreducible representations of $SU_q(2)$ can be expressed in terms of little q-Jacobi polynomials. Next Noumi and Mimachi [16] showed that the "spherical harmonics" on quantum homogeneous spaces of $SU_q(2)$ can be expressed in terms of big q-Jacobi polynomials. As a following step, the author [11] gave an interpretation of a two-parameter family of Askey–Wilson polynomials (including continuous q-Legendre polynomials) as (zonal) spherical functions on $SU_q(2)$. Here the notion of spherical function was generalized in the sense that bi-invariance with respect to the quantum subgroup $U(1)$ was replaced by "infinitesimal" left and right invariance with respect to twisted primitive elements of the corresponding quantized universal enveloping algebra $\mathcal{U}_q(sl(2,\mathbb{C}))$. Since this paper [11] was meant as a survey paper, the full results were only announced there, while a proof was sketched for the most simple case (parameter values $\sigma = \tau = 0$) corresponding to the continuous q-Legendre polynomials. Nevertheless, the paper [11] has already had some follow-ups by (i) the work of Koelink [8] (appearing as a companion paper to the present paper), which culminates in a quantum group derivation of the continuous q-Legendre case of the Rahman–Verma addition formula [19] for continuous q-ultraspherical polynomials, and (ii) an announcement by Noumi and Mimachi [17], [15], where they extend the author's result to the expression of all corresponding matrix elements (not just the left and right infinitesimally invariant ones) as Askey–Wilson polynomials.

It is the purpose of the present paper to give full proofs of the results announced in [11]. The contents are as follows. In §2 we give the preliminaries about q-hypergeometric functions and orthogonal polynomials, mainly referring to Askey

*Received by the editors August 8, 1990; accepted for publication (in revised form) September 8, 1992. The research for this paper was done at CWI, Amsterdam.

[†] University of Amsterdam, Faculty of Mathematics and Computer Science, Plantage Muidergracht 24, 1018 TV Amsterdam, the Netherlands.

and Wilson [5] and Gasper and Rahman [7]. In §3 we give preliminaries on the quantum group $SL_q(2,\mathbb{C})$. Section 4 introduces the (σ,τ)-spherical elements on $SU_q(2)$ and derives an explicit Fourier series for such elements belonging to irreducible representations. An important tool here is the explicit matrix of dual q-Krawtchouk polynomials for the basis change from the standard basis to a basis of eigenfunctions for an (almost) twisted primitive element. This last result (already announced in [11]) is also crucial in Koelink [8].

In §5 we prove that the elementary (σ,τ)-spherical matrix elements, when expressed as polynomials, satisfy the same second-order q-difference equation as the Askey–Wilson polynomials. This is done by use of the Casimir operator on the quantum group. This way of proving that our polynomials are Askey–Wilson polynomials is different from the proof we had in mind when writing [11]. There we worked with the explicit Fourier series and the knowledge obtained from the quantum group theory that we were dealing with orthogonal polynomials. Then the result could be derived by deriving the three-term recurrence relation. Section 5 contains also the expression of the Haar functional as an Askey–Wilson integral, when applied to (σ,τ)-spherical elements. Finally, in §6 we examine the limit cases as σ or $\tau \to \infty$. For the Askey–Wilson polynomials this means a limit transition to big or little q-Jacobi polynomials.

Notation. $\mathbb{Z}_+$ denotes the set of nonnegative integers.

2. Preliminaries on q-hypergeometric orthogonal polynomials.

Let $0 < q < 1$. Define *q-shifted factorials*

$$(a;q)_n := \prod_{k=0}^{n-1}(1-aq^k),$$
$$(a;q)_\infty := \lim_{n\to\infty}(a;q)_n\,,$$
$$(a_1,\ldots,a_r;q)_n := \prod_{j=1}^{r}(a_j;q)_n\,,$$

and the *q-hypergeometric series*

$$(2.1)\qquad {}_{s+1}\phi_s\left[\begin{matrix}a_1,\ldots,a_{s+1}\\ b_1,\ldots,b_s\end{matrix};q,z\right] := \sum_{k=0}^{\infty}\frac{(a_1,\ldots,a_{s+1};q)_k\,z^k}{(b_1,\ldots,b_s;q)_k\,(q;q)_k}\,.$$

Usually in this paper we will have the case of a *terminating* series in (2.1), i.e., $a_1 = q^{-n}$ ($n \in \mathbb{Z}_+$), so the series terminates after the term with $k = n$. Then we require that $b_1,\ldots,b_s \notin \{1,q^{-1},\ldots,q^{-n+1}\}$. See Gasper and Rahman [7, Ch. 1] for standard facts about q-hypergeometric series.

Askey-Wilson polynomials are defined by

(2.2)
$$p_n(\cos\theta;a,b,c,d\mid q) := a^{-n}\,(ab,ac,ad;q)_n\;{}_4\phi_3\left[\begin{matrix}q^{-n},q^{n-1}abcd,ae^{i\theta},ae^{-i\theta}\\ ab,ac,ad\end{matrix};q,q\right].$$

See Askey and Wilson [5, (1.15)]. They are symmetric in a,b,c,d (cf. [5, p. 6]). Sometimes it will be useful to write the ${}_4\phi_3$ factor in (2.2) as

$$(2.3)\qquad r_n(\cos\theta;a,b,c,d\mid q) := {}_4\phi_3\left[\begin{matrix}q^{-n},q^{n-1}abcd,ae^{i\theta},ae^{-i\theta}\\ ab,ac,ad\end{matrix};q,q\right].$$

The orthogonality properties are stated in [5, Thms. 2.2, 2.5].

PROPOSITION 2.1. *Assume a,b,c,d are real, or if complex, appear in conjugate pairs, and that $|a|,|b|,|c|,|d| \le 1$, but the pairwise products of a,b,c,d have absolute value less than one, then*

$$\frac{1}{2\pi}\int_0^\pi p_n(\cos\theta)\,p_m(\cos\theta)\,w(\cos\theta)\,d\theta = \delta_{m,n}\,h_n\,,$$

where

$$p_n(\cos\theta) = p_n(\cos\theta; a,b,c,d \mid q),$$

$$w(\cos\theta) = \frac{(e^{2i\theta}, e^{-2i\theta}; q)_\infty}{(ae^{i\theta}, ae^{-i\theta}, be^{i\theta}, be^{-i\theta}, ce^{i\theta}, ce^{-i\theta}, de^{i\theta}, de^{-i\theta}; q)_\infty}\,,$$

$$\frac{h_n}{h_0} = \frac{(1-q^{n-1}abcd)\,(q,ab,ac,ad,bc,bd,cd;q)_n}{(1-q^{2n-1}abcd)\,(abcd;q)_n}\,, \tag{2.4}$$

$$h_0 = \frac{(abcd;q)_\infty}{(q,ab,ac,ad,bc,bd,cd;q)_\infty}\,. \tag{2.5}$$

PROPOSITION 2.2. *Assume a,b,c,d are real, or if complex, appear in conjugate pairs, and that the pairwise products of a,b,c,d are not ≥ 1, then*

$$\frac{1}{2\pi}\int_0^\pi p_n(\cos\theta)\,p_m(\cos\theta)\,w(\cos\theta)\,d\theta + \sum_k p_n(x_k)\,p_m(x_k)\,w_k = \delta_{m,n}\,h_n\,, \tag{2.6}$$

where $p_n(\cos\theta)$, $w(\cos\theta)$ and h_n are as in Proposition 2.1, while the x_k are the points $(eq^k + e^{-1}q^{-k})/2$ with e any of the parameters $a, b, c,$ or d whose absolute value is larger than one; the sum is over the $k \in \mathbb{Z}_+$ with $|eq^k| > 1$ and w_k is $w_k(a;b,c,d)$ as defined by [5, (2.10)] *when $x_k = (aq^k + a^{-1}q^{-k})/2$. (Be aware that $(1-aq^{2k})/(1-a)$ should be replaced by $(1-a^2q^{2k})/(1-a^2)$ in* [5, (2.10)].)

With notation as in Proposition 2.2 let $dm = dm_{a,b,c,d;q}$ be the normalized orthogonality measure for the Askey–Wilson polynomials. So, for any polynomial p,

$$\int_{-\infty}^{\infty} p(x)dm(x) = \frac{1}{h_0}\left\{\frac{1}{2\pi}\int_{-1}^{1} p(x)\,w(x)\,\frac{dx}{(1-x^2)^{\frac12}} + \sum_k p(x_k)\,w_k\right\}. \tag{2.7}$$

By [5, (5.7)–(5.9)] the Askey–Wilson polynomials, written as

$$R_n(e^{i\theta}) := {}_4\phi_3\left[\begin{matrix} q^{-n}, q^{n-1}abcd, ae^{i\theta}, ae^{-i\theta} \\ ab, ac, ad \end{matrix}; q, q\right],$$

are eigenfunctions of a second-order q-difference operator:

$$\begin{aligned} &A(-\theta)\,(R_n(q^{-1}e^{i\theta}) - R_n(e^{i\theta})) + A(\theta)(R_n(qe^{i\theta}) - R_n(e^{i\theta})) \\ &\quad = -(1-q^{-n})\,(1-q^{n-1}abcd)\,R_n(e^{i\theta}), \end{aligned} \tag{2.8}$$

where

$$A(\theta) := \frac{(1-ae^{i\theta})\,(1-be^{i\theta})\,(1-ce^{i\theta})\,(1-de^{i\theta})}{(1-e^{2i\theta})\,(1-qe^{2i\theta})}\,.$$

If $f(e^{i\theta})$ is a polynomial of degree $\le n$ in $\cos\theta$ and if (2.8) with R_n replaced by f is valid, then f will be a constant multiple of R_n.

We will need some special Askey–Wilson polynomials which happen to have simple explicit Fourier-cosine expansions: Chebyshev polynomials of the first kind

$$p_n(\cos\theta; 1, -1, q^{\frac{1}{2}}, -q^{\frac{1}{2}} \mid q) = (2 - \delta_{n,0})\,(q^n; q)_n \cos(n\theta) \tag{2.9}$$

(cf. [5, (4.25)]) and *continuous* q-Legendre polynomials

$$\begin{aligned} &p_n(\cos\theta; q^{\frac{1}{2}}, -q^{\frac{1}{2}}, q^{\frac{1}{2}}, -q^{\frac{1}{2}} \mid q) \\ &\quad = (-q;q)_n^2\,(q;q)_n \sum_{k=0}^{n} \frac{(q;q^2)_k\,(q;q^2)_{n-k}}{(q^2;q^2)_k\,(q^2;q^2)_{n-k}}\, e^{i(n-2k)\theta} \end{aligned}$$

(cf. [3, (3.1)] together with [5, (4.2), (4.20)]).

LEMMA 2.3. *The connection coefficients* $c_{k,n}$ *in*

$$p_n(x; q^{\frac{1}{2}}a, \beta, qa^2/\beta, q^{\frac{1}{2}}a \mid q) = \sum_{k=0}^{n} c_{k,n}\, p_k(x; a, -a, -q^{\frac{1}{2}}a, q^{\frac{1}{2}}a \mid q) \tag{2.10}$$

can be explicitly written as

$$\begin{aligned} c_{k,n} &= \frac{(qa^2, q; q)_n\,(q^{n+1}a^4; q)_k}{(qa^2, q, q^k a^4; q)_k\,(q;q)_{n-k}} \left(\frac{-q^{\frac{1}{2}}}{a^2\beta}\right)^{n-k} q^{(n-k)^2/2} \\ &\quad \times \left\{ {}_2\phi_1\left[\begin{matrix} q^{-n+k}, q^{-n}a^2 \\ q^{k+1}a^2 \end{matrix}; q, q^{n+\frac{1}{2}}a\beta\right]\right\}^2 . \end{aligned} \tag{2.11}$$

Proof. Askey and Wilson [5, (6.1), (6.2)] gave the connection coefficients between two families of Askey–Wilson polynomials with one common parameter. Their expression involved a terminating balanced ${}_5\phi_4$ of argument q. With our special choice of parameters in (2.10) this ${}_5\phi_4$ has the form occurring in the q-Clausen formula as given by formulas (2.16) (first identity) and (2.4) (with $\alpha = q^{-n}$) in Gasper and Rahman [6]. (Note that both the left- and right-hand side of this version of the q-Clausen formula are written differently from the usual formulation [6, (1.6)].) Substitution of this q-Clausen formula yields (2.11). □

Now substitute $a = 1$ in (2.10), (2.11) and apply (2.9). Then

$$p_n(\cos\theta; q^{\frac{1}{2}}, \beta, q/\beta, q^{\frac{1}{2}} \mid q) = \sum_{k=-n}^{n} c_{|k|,n}\,(q^{|k|}; q)_{|k|}\, e^{ik\theta},$$

with $c_{k,n}$ given by (2.11) for $a = 1$. When we switch to base q^2 and put $\beta = -q^{2\sigma+1}$ we obtain

$$p_n(\cos\theta; q, -q^{2\sigma+1}, -q^{-2\sigma+1}, q \mid q^2) = \sum_{k=-n}^{n} c_{|k|,n}\,(q^{2|k|}; q^2)_{|k|}\, e^{ik\theta} \tag{2.12}$$

with

$$\begin{aligned} (q^{2k}; q^2)_k\, c_{k,n} &= \frac{(q^2;q^2)_n^2\,(q^{2n+2};q^2)_k\, q^{(n-k)(n-k+2\sigma)}}{(q^2;q^2)_k^2\,(q^2;q^2)_{n-k}} \\ &\quad \times \left\{ {}_2\phi_1\left[\begin{matrix} q^{-2n+2k}, q^{-2n} \\ q^{2k+2} \end{matrix}; q^2, -q^{2n-2\sigma+2}\right]\right\}^2 . \end{aligned} \tag{2.13}$$

Finally, by Jackson's transformation formula [7, (III.7)], (2.13) can be rewritten as

$$\begin{aligned}(q^{2k};q^2)_k\, c_{k,n} =& \frac{(q^2;q^2)_{2n}^2\, q^{(n-k)(n-k+2\sigma)}}{(q^2;q^2)_n\,(q^2;q^2)_{n+k}\,(q^2;q^2)_{n-k}} \\ &\times \left\{ {}_3\phi_2\left[\begin{matrix} q^{-2n+2k}, q^{-2n}, -q^{-2n-2\sigma} \\ q^{-4n}, 0 \end{matrix}; q^2, q^2\right]\right\}^2 .\end{aligned} \tag{2.14}$$

Next we define *little q-Jacobi polynomials*

$$p_n(x;a,b;q) := {}_2\phi_1(q^{-n}, abq^{n+1}; aq; q, qx) \tag{2.15}$$

(cf. Andrews and Askey [1]) and *big q-Jacobi polynomials*

$$P_n^{(\alpha,\beta)}(x;c,d;q) = {}_3\phi_2\left[\begin{matrix} q^{-n}, q^{n+\alpha+\beta+1}, xq^{\alpha+1}/c \\ q^{\alpha+1}, -q^{\alpha+1}d/c \end{matrix}; q, q\right] \tag{2.16}$$

(cf. Andrews and Askey [2]). In this last reference some different normalization is suggested for the big q-Jacobi polynomials, but the authors are not very definite about it. Here we follow the normalization used by Noumi and Mimachi [16]. Little and big q-Jacobi polynomials are orthogonal with respect to discrete measures.

We define *dual q-Krawtchouk polynomials* by

$$R_n(q^{-x} - q^{x-N-c}; q^c, N \mid q) := {}_3\phi_2(q^{-n}, q^{-x}, -q^{x-N-c}; 0, q^{-N}; q, q). \tag{2.17}$$

These are special q-Racah polynomials and satisfy the orthogonality relations

$$\begin{aligned}&\frac{1}{(-q^c;q)_N}\sum_{x=0}^{N}(R_nR_m)(q^{-x} - q^{x-N-c}; q^c, N \mid q) \\ &\times \frac{(1+q^{2x-N-c})(-q^{-N-c}, q^{-N}; q)_x}{(1+q^{-N-c})(q, -q^{-c+1}; q)_x(-q^{x-2N-c})^x} = \delta_{nm}\,\frac{(q;q)_n}{(q^{-N};q)_n}\,(-q^{-N-c})^n,\end{aligned} \tag{2.18}$$

where $n, m = 0, \ldots, N$. See Askey and Wilson [4] and Stanton [20]. They satisfy (see [20]) the three-term recurrence relation

$$\begin{aligned} y\, R_n(y; q^c, N \mid q) =& (1 - q^{n-N})\, R_{n+1}(y; q^c, N \mid q) \\ &+ (q^{-N} - q^{-N-c})\, q^n\, R_n(y; q^c, N \mid q) \\ &- (1 - q^n)\, q^{-N-c}\, R_{n-1}(y; q^c, N \mid q). \end{aligned} \tag{2.19}$$

3. Preliminaries on the quantum $SL(2,\mathbb{C})$ group. The reader may use the author's survey [11] and the references given there for further reading in connection with this section. Fix $q \in (0,1)$. Let $\mathcal{A}_q$ be the complex associative algebra with unit 1, generators $\alpha, \beta, \gamma, \delta$, and relations

$$\begin{aligned} &\alpha\beta = q\beta\alpha, \quad \alpha\gamma = q\gamma\alpha, \quad \beta\delta = q\delta\beta, \quad \gamma\delta = q\delta\gamma, \quad \beta\gamma = \gamma\beta, \\ &\alpha\delta - q\beta\gamma = \delta\alpha - q^{-1}\beta\gamma = 1. \end{aligned} \tag{3.1}$$

It turns out that $\mathcal{A}_q$ becomes a Hopf algebra over $\mathbb{C}$ under the following actions of the comultiplication $\Delta\colon \mathcal{A}_q \to \mathcal{A}_q \otimes \mathcal{A}_q$, counit $\varepsilon\colon \mathcal{A}_q \to \mathbb{C}$ (unital multiplicative linear mappings), and antipode $S: \mathcal{A}_q \to \mathcal{A}_q$ (unital antimultiplicative linear mapping) on the generators

$$\Delta\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix} = \begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix} \otimes \begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix},$$
$$\varepsilon\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix} = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}, \quad S\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix} = \begin{pmatrix}\delta & -q^{-1}\beta\\ -q\gamma & \alpha\end{pmatrix},$$

where the formula for Δ has to be interpreted in the sense of matrix multiplication: $\Delta(\alpha) = \alpha \otimes \alpha + \beta \otimes \gamma$, etc.

A *Hopf $*$-algebra* is a Hopf algebra $\mathcal{A}$ over $\mathbb{C}$ with an involution $a \mapsto a^*$ such that $\mathcal{A}$ becomes a unital $*$-algebra and $\Delta: \mathcal{A} \to \mathcal{A} \otimes \mathcal{A}$ and $\varepsilon: \mathcal{A} \to \mathbb{C}$ are $*$-homomorphisms. Then it can be shown that S is invertible and that $S \circ * \circ S \circ * = \mathrm{id}$. We can make $\mathcal{A}_q$ into a Hopf $*$-algebra by taking for the involution the unital antimultiplicative antilinear mapping $a \mapsto a^*$ such that

$$\begin{pmatrix}\alpha^* & \beta^*\\ \gamma^* & \delta^*\end{pmatrix} = \begin{pmatrix}\delta & -q\gamma\\ -q^{-1}\beta & \alpha\end{pmatrix}.$$

Let $\mathcal{U}_q$ be the complex associative algebra with unit 1, generators A, B, C, D, and relations

$$AD = DA = 1, \quad AB = qBA, \quad AC = q^{-1}CA, \quad BC - CB = \frac{A^2 - D^2}{q - q^{-1}}. \tag{3.2}$$

We can make $\mathcal{U}_q$ into a Hopf $*$-algebra with comultiplication $\Delta: \mathcal{U}_q \to \mathcal{U}_q \otimes \mathcal{U}_q$, counit $\varepsilon: \mathcal{U}_q \to \mathbb{C}$, antipode $S: \mathcal{U}_q \to \mathcal{U}_q$, and involution $*: \mathcal{U}_q \to \mathcal{U}_q$ by requiring that

$$\begin{aligned}
&\Delta(A) = A \otimes A, \quad \Delta(D) = D \otimes D,\\
&\Delta(B) = A \otimes B + B \otimes D, \quad \Delta(C) = A \otimes C + C \otimes D,\\
&\varepsilon(A) = \varepsilon(D) = 1, \quad \varepsilon(B) = \varepsilon(C) = 0,\\
&S(A) = D, \quad S(D) = A, \quad S(B) = -q^{-1}B, \quad S(C) = -qC, && (3.3)\\
&\quad A^* = A, \quad D^* = D, \quad B^* = C, \quad C^* = B. && (3.4)
\end{aligned}$$

Two Hopf algebras $\mathcal{U}, \mathcal{A}$ are said to be *in duality* if there is a doubly nondegenerate bilinear form $(u, a) \mapsto \langle u, a\rangle: \mathcal{U} \times \mathcal{A} \to \mathbb{C}$ such that, for $u, v \in \mathcal{U}$, $a, b \in \mathcal{A}$,

$$\begin{aligned}
&\langle \Delta(u), a \otimes b\rangle = \langle u, ab\rangle, \quad \langle u \otimes v, \Delta(a)\rangle = \langle uv, a\rangle,\\
&\quad \langle 1_{\mathcal{U}}, a\rangle = \varepsilon_{\mathcal{A}}(a), \quad \langle u, 1_{\mathcal{A}}\rangle = \varepsilon_{\mathcal{U}}(u), \quad \langle S(u), a\rangle = \langle a, S(u)\rangle.
\end{aligned} \tag{3.5}$$

If $\mathcal{U}, \mathcal{A}$ are moreover Hopf $*$-algebras, then they are said to be *Hopf $*$-algebras in duality* if the above pairing satisfies in addition that

$$\langle u^*, a\rangle = \overline{\langle u, (S(a))^*\rangle}. \tag{3.6}$$

Instead of $\langle u, a\rangle$ we will also write $u(a)$ or $a(u)$.

It can be shown that $\mathcal{U}_q$ and $\mathcal{A}_q$ become Hopf $*$-algebras in duality with the following pairing between the generators:

$$\left\langle A, \begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}\right\rangle = \begin{pmatrix}q^{\frac{1}{2}} & 0\\ 0 & q^{-\frac{1}{2}}\end{pmatrix}, \quad \left\langle D, \begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}\right\rangle = \begin{pmatrix}q^{-\frac{1}{2}} & 0\\ 0 & q^{\frac{1}{2}}\end{pmatrix},$$
$$\left\langle B, \begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}\right\rangle = \begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}, \quad \left\langle C, \begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}\right\rangle = \begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix}.$$

The pairing between products of generators then follows by the rules (3.5).

The element

$$\Omega := \frac{q^{-1}A^2 + qD^2 - 2}{(q^{-1} - q)^2} + BC = \Omega^* \tag{3.7}$$

is a *Casimir element* of $\mathcal{U}_q$: it commutes with any $X \in \mathcal{U}_q$.

In a Hopf algebra $\mathcal{U}$ an element u is called *group-like* if $\Delta(u) = u \otimes u$, *primitive* if $\Delta(u) = 1 \otimes u + u \otimes 1$, and *twisted primitive* (with respect to a group-like element g) if $\Delta(u) = g \otimes u + u \otimes S(g)$. In $\mathcal{U}_q$ the group-like elements are all elements A^n $(n \in \mathbb{Z})$ and (cf. Masuda et al. [12]).

LEMMA 3.1. *The twisted primitive elements with respect to A are the elements X in the linear span of $A - D$, B, and C. They satisfy*

$$\Delta(X) = A \otimes X + X \otimes D.$$

For $t \neq 1$ the twisted primitive elements with respect to A^t are the constant multiples of $A^t - A^{-t}$.

Let $\mathcal{U}$ and $\mathcal{A}$ be Hopf algebras in duality. For $u \in \mathcal{U}$ and $a \in \mathcal{A}$ define elements $u.a$ and $a.u$ of $\mathcal{A}$ by

$$u.a := (\mathrm{id} \otimes u)(\Delta(a)), \qquad a.u := (u \otimes \mathrm{id})(\Delta(a)). \tag{3.8}$$

Hence, if $v \in \mathcal{U}$,

$$(u.a)(v) = a(vu), \qquad (a.u)(v) = a(uv).$$

The operations defined in (3.8) are left, respectively, right algebra actions of $\mathcal{U}_q$ on $\mathcal{A}_q$:

$$(uv).a = u.(v.a), \quad a.(uv) = (a.u).v. \tag{3.9}$$

If $\Delta(u) = \sum_{(u)} u_{(1)} \otimes u_{(2)}$ $(u \in \mathcal{U})$ and $a, b \in \mathcal{A}$, then

$$u.(ab) = \sum_{(u)} (u_{(1)}.a)\,(u_{(2)}.b), \qquad (ab).u = \sum_{(u)} (a.u_{(1)})\,(b.u_{(2)}). \tag{3.10}$$

Furthermore, if $u \in \mathcal{U}$, $a \in \mathcal{A}$, then

$$(\mathrm{id} \otimes u.)(\Delta(a)) = \Delta(u.a); \tag{3.11}$$

and if, moreover, $v \in \mathcal{U}$, then

$$\langle v, u.a^* \rangle = \overline{\langle S(v)^*, S(u)^*.a \rangle}. \tag{3.12}$$

We call an element $a \in \mathcal{A}$ *left* (*right*) *invariant* with respect to an element $u \in \mathcal{U}$ if $u.a = \varepsilon(u)a$, respectively, $a.u = \varepsilon(u)a$. Note that the unit 1 of $\mathcal{A}$ is bi-invariant with respect to all $u \in \mathcal{U}$. If u is twisted primitive, then $\varepsilon(u) = 0$, and

$$\begin{aligned} u.a = 0 \quad &\text{and} \quad b.u = 0 \Longrightarrow u.(ab) = 0, \\ a.u = 0 \quad &\text{and} \quad b.u = 0 \Longrightarrow (ab).u = 0. \end{aligned}$$

LEMMA 3.2. *The left (or right) invariant elements of $\mathcal{A}$ with respect to some twisted primitive element of $\mathcal{U}$ form a unital subalgebra of $\mathcal{A}$. In particular, if $X \in \operatorname{Span}\{A-D, B, C\}$, then the set of all $a \in \mathcal{A}_q$ satisfying $X.a = 0$ (respectively $a.X = 0$) forms a unital subalgebra of $\mathcal{A}_q$.*

Let $\mathcal{U}$ and $\mathcal{A}$ be Hopf algebras in duality. A *matrix corepresentation* of $\mathcal{A}$ is a square matrix $t = (t_{i,j})$ of elements of $\mathcal{A}$ such that

$$\Delta(t_{i,j}) = \sum_k t_{i,k} \otimes t_{k,j}, \qquad \varepsilon(t_{i,j}) = \delta_{i,j}. \tag{3.13}$$

To a matrix corepresentation t of $\mathcal{A}$ corresponds a matrix representation of $\mathcal{U}$, also denoted by t and defined by

$$(t(u))_{i,j} := t_{i,j}(u) = \langle u, t_{i,j} \rangle, \qquad u \in \mathcal{U}.$$

The matrix entries of a corepresentation of $\mathcal{A}$ (elements of $\mathcal{A}$) are completely determined by the matrix entries of the corresponding representation of $\mathcal{U}$ (linear functionals on $\mathcal{U}$). A matrix corepresentation t of $\mathcal{A}$ is called *unitary* if $t^*_{i,j} = S(t_{j,i})$ and a representation t of $\mathcal{U}$ is called a *$*$-representation* if $t_{i,j}(u^*) = \overline{t_{j,i}(u)}$ $(u \in \mathcal{U})$. Note that a matrix corepresentation of $\mathcal{A}$ is unitary if and only if the corresponding matrix representation of $\mathcal{U}$ is a $*$-representation.

Up to equivalence, there is for each finite dimension precisely one irreducible matrix corepresentation of $\mathcal{A}_q$, which can be chosen to be unitary. The corresponding irreducible $*$-representation of $\mathcal{U}_q$ is realized as a representation $t^l = (t^l_{i,j})_{i,j=-l,-l+1,\ldots,l}$ $(l \in \frac{1}{2}\mathbb{Z}_+)$ on a $(2l+1)$-dimensional vector space with $\{e^l_n\}_{n=-l,-l+1,\ldots,l}$ as an orthonormal basis such that

$$\begin{aligned} &t^l(A)\, e^l_n = q^{-n}\, e^l_n, \qquad t^l(D)\, e^l_n = q^n\, e^l_n, \\ &t^l(B)\, e^l_n = \frac{(q^{-l+n-1} - q^{l-n+1})^{\frac{1}{2}}\, (q^{-l-n} - q^{l+n})^{\frac{1}{2}}}{q^{-1} - q}\, e^l_{n-1}, \\ &t^l(C)\, e^l_n = \frac{(q^{-l+n} - q^{l-n})^{\frac{1}{2}}\, (q^{-l-n-1} - q^{l+n+1})^{\frac{1}{2}}}{q^{-1} - q}\, e^l_{n+1}, \end{aligned} \tag{3.14}$$

with the convention that e^l_{-l-1} and e^l_{l+1} are zero. The $t^l_{i,j}$, being elements of $\mathcal{A}_q$, can be expressed in terms of the generators by expressions involving little q-Jacobi polynomials. The lowest dimensional cases are particularly simple:

$$t^0 = (t^0_{0,0}) = (1), \qquad t^{\frac{1}{2}} = \begin{pmatrix} t^{\frac{1}{2}}_{\frac{1}{2},\frac{1}{2}} & t^{\frac{1}{2}}_{\frac{1}{2},-\frac{1}{2}} \\ t^{\frac{1}{2}}_{-\frac{1}{2},\frac{1}{2}} & t^{\frac{1}{2}}_{-\frac{1}{2},-\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} \delta & \gamma \\ \beta & \alpha \end{pmatrix},$$

(3.15)

$$t^1 = \begin{pmatrix} t^1_{1,1} & t^1_{1,0} & t^1_{1,-1} \\ t^1_{0,1} & t^1_{0,0} & t^1_{0,-1} \\ t^1_{-1,1} & t^1_{-1,0} & t^1_{-1,-1} \end{pmatrix} = \begin{pmatrix} \delta^2 & (1+q^2)^{\frac{1}{2}}\delta\gamma & \gamma^2 \\ (1+q^2)^{\frac{1}{2}}\delta\beta & 1 + (q+q^{-1})\beta\gamma & (1+q^2)^{\frac{1}{2}}\gamma\alpha \\ \beta^2 & (1+q^2)^{\frac{1}{2}}\beta\alpha & \alpha^2 \end{pmatrix}.$$

Define

$$\mathcal{A}^l_q := \operatorname{Span}\{t^l_{i,j} \mid i,j = -l, -l+1, \ldots, l\}. \tag{3.16}$$

PROPOSITION 3.3. *The $t^l_{i,j}$ $(i,j=-l,-l+1,\dots,l)$ form a basis of $\mathcal{A}^l_q$ and*

$$\mathcal{A}_q = \bigoplus\nolimits_{l\in\frac12\mathbb{Z}_+} \mathcal{A}^l_q\,. \tag{3.17}$$

Let $a := \sum_{i,j=-l}^{l} \gamma_{i,j}\, t^l_{i,j} \in \mathcal{A}^l_q$ $(\gamma_{i,j}\in\mathbb{C})$. Let $X\in\mathcal{U}_q$. It follows from (3.8) and (3.13) that

$$X.a = \sum\nolimits_{i,k} \left(\sum\nolimits_j t^l_{k,j}(X)\,\gamma_{i,j}\right) t^l_{i,k}, \tag{3.18}$$

$$a.X = \sum\nolimits_{k,j} \left(\sum\nolimits_i t^l_{i,k}(X)\,\gamma_{i,j}\right) t^l_{k,j}. \tag{3.19}$$

LEMMA 3.4. *With a and X as above we have the following:*

(i) $X.a = 0 \iff t^l(X)\left(\sum_{j=-l}^{l}\gamma_{ij}\, e^l_j\right) = 0$ *for all* i;

(ii) $a.X = 0 \iff t^l(X^*)\left(\sum_{i=-l}^{l}\overline{\gamma_{ij}}\, e^l_i\right) = 0$ *for all* j.

Proof. (i) By (3.18) and Proposition 3.3, $X.a=0 \iff \sum_j t^l_{k,j}(X)\,\gamma_{i,j} = 0$ for all $i,k \iff t^l(X)\left(\sum_{j=-l}^{l}\gamma_{ij}\,e^l_j\right)=0$ for all i.

(ii) By (3.19), $a.X = 0 \iff \sum_i t^l_{i,k}(X)\,\gamma_{i,j} = 0$ for all $k,j \iff \sum_i t^l_{k,i}(X^*)\,\overline{\gamma_{i,j}} = 0$ for all $k,j \iff t^l(X^*)\left(\sum_i \overline{\gamma_{i,j}}\, e^l_i\right) = 0$ for all j, where we used that t^l is a $*$-representation of $\mathcal{U}_q$. □

For the Casimir element Ω (cf. (3.7)) we compute from (3.14) that

$$t^l(\Omega) = \left(\frac{q^{-l-\frac12} - q^{l+\frac12}}{q^{-1}-q}\right)^2 \mathrm{id}\,.$$

Hence, by (3.18), (3.19),

$$\Omega.a = \left(\frac{q^{-l-\frac12} - q^{l+\frac12}}{q^{-1}-q}\right)^2 a = a.\Omega \quad \text{if } a\in\mathcal{A}^l_q\,. \tag{3.20}$$

The *tensor product* $t^l\otimes t^{l'}$ is defined as the matrix corepresentation of $\mathcal{A}_q$ with matrix entries

$$(t^l\otimes t^{l'})_{i,i';j,j'} := t^l_{i,j}\, t^{l'}_{i',j'} \qquad (i,j=-l,-l+1,\dots,l;\ i',j'=-l',-l'+1,\dots,l').$$

PROPOSITION 3.5. *$t^l\otimes t^{l'}$ is equivalent to the direct sum of the corepresentations t^k $(k=l+l',l+l'-1,\dots,|l-l'|)$.*

There is a unique linear functional $h\colon \mathcal{A}_q\to\mathbb{C}$, called the *Haar functional* on $\mathcal{A}_q$, with the properties

(i) $h(1)=1$;

(ii) $h(aa^*)\ge 0$ for all $a\in\mathcal{A}_q$;

(iii) $(h\otimes\mathrm{id})(\Delta(a)) = h(a)1 = (\mathrm{id}\otimes h)(\Delta(a))$.

Then $h(aa^*)>0$ if $a>0$. It can also be shown that

$$h((t^{l'}_{i',j'})^*\, t^l_{i,j}) = \delta_{l,l'}\,\delta_{i,i'}\,\delta_{j,j'}\, q^{2(l-i)}\,\frac{1-q^2}{1-q^{2(2l+1)}}\,. \tag{3.21}$$

For $\theta\in\mathbb{C}$ let $\pi^1_\theta\colon\mathcal{A}_q\to\mathbb{C}$ be the unital algebra homomorphism (one-dimensional representation of $\mathcal{A}_q$) defined by

$$\pi^1_\theta\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix} := \begin{pmatrix} e^{i\theta} & 0\\ 0 & e^{-i\theta}\end{pmatrix}. \tag{3.22}$$

In particular, if $\theta \in \mathbb{R}$, π_θ^1 is a $*$-representation of $\mathcal{A}_q$. We have

$$\pi_\theta^1(t_{n,m}^l) = e^{-2in\theta}\,\delta_{n,m}. \tag{3.23}$$

4. (σ,τ)-spherical elements. Define, for $\sigma \in \mathbb{R}$,

$$X_\sigma := i\,q^{\frac12}\,B - i\,q^{-\frac12}\,C - \frac{q^{-\sigma}-q^{\sigma}}{q^{-1}-q}\,(A-D). \tag{4.1}$$

Then

$$X_\sigma^* = i\,q^{-\frac12}\,B - i\,q^{\frac12}\,C - \frac{q^{-\sigma}-q^{\sigma}}{q^{-1}-q}\,(A-D) = -S(X_\sigma) \tag{4.2}$$

by (3.3), (3.4), and X_σ is twisted primitive with respect to A (cf. Lemma 3.1):

$$\Delta(X_\sigma) = A\otimes X_\sigma + X_\sigma \otimes D.$$

Define also the twisted primitive element

$$X_\infty := D - A = \lim_{\sigma\to\infty}(q^{-1}-q)\,q^{\sigma}\,X_\sigma = \lim_{\sigma\to-\infty}(q-q^{-1})\,q^{-\sigma}\,X_\sigma.$$

Left or right invariance of an element of $\mathcal{A}_q$ with respect to X_∞ is the same as left or right invariance with respect to the diagonal quantum subgroup of $SL_q(2,\mathbb{C})$.

We will call an element $a \in \mathcal{A}_q$ *(σ,τ)-spherical* if a is left invariant with respect to X_σ and right invariant with respect to X_τ:

$$X_\sigma.a = 0 \quad\text{and}\quad a.X_\tau = 0.$$

The nonzero (σ,τ)-spherical elements in the subspaces $\mathcal{A}_q^l$ (defined by (3.16)) will be called *elementary (σ,τ)-spherical.* It follows from Lemma 3.2 that the (σ,τ)-spherical elements form a subalgebra with 1 of $\mathcal{A}_q$. Since the subspaces $\mathcal{A}_q^l$ are invariant under left and right action of $\mathcal{U}_q$ (cf. (3.18), (3.19)), it follows from the direct sum decomposition (3.17) that any (σ,τ)-spherical element in $\mathcal{A}_q$ will be a sum of elementary (σ,τ)-spherical elements.

From now on assume that σ and τ are finite. From (3.14) we obtain the following.

LEMMA 4.1. *Let $l \in \frac12\mathbb{Z}_+$, $a_{-l}, a_{-l+1}, \dots, a_l \in \mathbb{C}$. Let $v := \sum_{m=-l}^{l} a_m\, e_m^l$. Then*

$$t^l\left(iq^{\pm\frac12}B - iq^{\mp\frac12}C - \frac{q^{-\sigma}-q^{\sigma}}{q^{-1}-q}\,(A-D)\right)v = 0 \tag{4.3}$$

if and only if
(4.4)

$$i\,q^{\pm\frac12}\,(q^{-l+m}-q^{l-m})^{\frac12}\,(q^{-l-m-1}-q^{l+m+1})^{\frac12}\,a_{m+1} - i\,q^{\mp\frac12}\,(q^{-l+m-1}-q^{l-m+1})^{\frac12}$$
$$\times\,(q^{-l-m}-q^{l+m})^{\frac12}\,a_{m-1} - (q^{-\sigma}-q^{\sigma})\,(q^{-m}-q^{m})\,a_m = 0, \quad m = -l, -l+1, \dots, l,$$

with the convention that $a_{-l-1} = 0 = a_{l+1}$.

By Lemma 4.1 we easily find the general solution to (4.3) for low l: if $l = 0$, then $v = \text{const.}$; if $l = \frac12$, then $v = 0$; if $l = 1$, then

$$v = \text{const.}\left(-i\,q^{\mp\frac12}\,e_1^1 + \frac{q^{-\sigma}-q^{\sigma}}{(q^{-1}+q)^{\frac12}}\,e_0^1 - iq^{\pm\frac12}\,e_{-1}^1\right). \tag{4.5}$$

Also, if $\sigma = 0$, then the coefficient of a_m in (4.4) vanishes, so (4.4) becomes a two-term recurrence relation with solution $v = 0$ if $l \in \mathbb{Z}_+ + \frac{1}{2}$, and

$$v = \text{const.} \sum_{m=-l,-l+2,\ldots,l} q^{\mp\frac{1}{2}m} \left(\frac{(q^2;q^4)_{(l-m)/2}\,(q^2;q^4)_{(l+m)/2}}{(q^4;q^4)_{(l-m)/2}\,(q^4;q^4)_{(l+m)/2}} \right)^{\frac{1}{2}} e^l_m$$

if $l \in \mathbb{Z}_+$.

From Lemma 3.4 we derive the following.

LEMMA 4.2. *Let, for some $l \in \frac{1}{2}\mathbb{Z}_+$ and $\sigma, \tau \in \mathbb{R}$, $t^l(X_\sigma)$ and $t^l(X^*_\tau)$ have both one-dimensional zero-space spanned by $\sum_{m=-l}^{l} a_m e^l_m$ and $\sum_{m=-l}^{l} b_m e^l_m$, respectively. Then the (σ,τ)-spherical elements in $\mathcal{A}^l_q$ form a one-dimensional subspace spanned by*

$$\sum_{i,j=-l}^{l} \overline{b_i}\, a_j\, t^l_{i,j}\,. \tag{4.6}$$

In view of (3.15), (4.5), and (4.6), the (σ,τ)-spherical elements in $\mathcal{A}^1_q$ are just the constant multiples of

$$2\rho_{\sigma,\tau} + \frac{(q^{-\sigma} - q^{\sigma})(q^{-\tau} - q^{\tau})}{q^{-1} + q}\,, \tag{4.7}$$

where

$$\begin{aligned} \rho_{\sigma,\tau} :=& \tfrac{1}{2}\big(\alpha^2 + \delta^2 + q\gamma^2 + q^{-1}\beta^2 + i(q^{-\sigma} - q^{\sigma})(q\delta\gamma + \beta\alpha) \\ &- i(q^{-\tau} - q^{\tau})(\delta\beta + q\gamma\alpha) + (q^{-\sigma} - q^{\sigma})(q^{-\tau} - q^{\tau})\beta\gamma\big). \end{aligned} \tag{4.8}$$

Note that

$$\rho^*_{\sigma,\tau} = \rho_{\sigma,\tau}\,.$$

In order to compute the null space of $t^l(X_\sigma)$ in general, we consider the more general problem of finding the eigenvectors of $t^l(-DX_\sigma)$. Note that

$$(DX_\sigma)^* = DX_\sigma;$$

hence $t^l(-DX_\sigma)$ is selfadjoint. Clearly, $t^l(X_\sigma)$ and $t^l(-DX_\sigma)$ have the same zero space.

Let $\lambda \in \mathbb{R}$, $v = \sum_{m=-l}^{l} a_m\, e^l_m$. Then

$$t^l(-DX_\sigma)\, v = \lambda\, v$$

if and only if

(4.9)

$$\begin{aligned} -i\, q^{\frac{1}{2}}\, (q^{-l+m} - q^{l-m})^{\frac{1}{2}}\, (q^{-l-m-1} - q^{l+m+1})^{\frac{1}{2}}\, a_{m+1} + i\, q^{-\frac{1}{2}}\, (q^{-l+m-1} - q^{l-m+1})^{\frac{1}{2}} \\ \times (q^{-l-m} - q^{l+m})^{\frac{1}{2}}\, a_{m-1} + (q^{-\sigma} - q^{\sigma})\,(q^{-m} - q^{m})\, a_m = q^{-m}\, \lambda\, (q^{-1} - q)\, a_m \end{aligned}$$

for $m = -l, -l+1, \ldots, l$. Put

$$\begin{aligned} R_n &:= i^n\, q^{-\frac{1}{2}n(n-1)}\, q^{-n\sigma}\, (q^2;q^2)_n^{\frac{1}{2}}\, (q^{4l};q^{-2})_n^{-\frac{1}{2}}\, a_{-l+n}\,, \\ x &:= q^{-2l-\sigma}\, ((q^{-1} - q)\lambda + q^{\sigma} - q^{-\sigma}). \end{aligned}$$

Then (4.9) can be rewritten as

(4.10)
$$(1-q^{2n-4l})\,R_{n+1}-(1-q^{2n})\,q^{-4l-2\sigma}\,R_{n-1}+(q^{-4l}-q^{-4l-2\sigma})\,q^{2n}\,R_n=x\,R_n\,,\qquad n=0,1,\ldots,2l,$$

with the convention that $R_{-1}=0=R_{2l+1}$. We recognize (4.10) as the three-term recurrence relation (2.19) for the dual q-Krawtchouk polynomials (2.17). Thus the possible eigenvalues in (4.10) are

(4.11)
$$x_j:=q^{-2j-2l}-q^{2j-2l-2\sigma},\qquad j=-l,-l+1,\ldots,l,$$

and the corresponding eigenvectors, up to a constant factor, are given by

$$R_n=R_n(x_j;q^{2\sigma},2l\mid q^2).$$

When we translate this back to (4.9) we obtain the following.

THEOREM 4.3. *$t^l(-DX_\sigma)$ has simple spectrum consisting of eigenvalues*

$$\lambda_j:=\frac{-q^{2j-\sigma}+q^{\sigma-2j}-q^{\sigma}+q^{-\sigma}}{q^{-1}-q},\qquad j=-l,-l+1,\ldots,l,$$

with corresponding eigenvectors

$$\text{const.}\sum_{n=0}^{2l}i^{-n}\,q^{n\sigma}\,q^{n(n-1)/2}\,(q^2;q^2)_n^{-\frac12}\,(q^{4l};q^{-2})_n^{\frac12}\,R_n(x_j;q^{2\sigma},2l\mid q^2)\,e^l_{n-l}\,,$$

where x_j is given by (4.11).

Similarly, in order to compute the null space of $t^l(X_\sigma^*)$ in general, we consider the more general problem of finding the eigenvectors of $t^l(AX_\sigma^*)$. Note that

$$(AX_\sigma^*)^*=AX_\sigma^*;$$

so $t^l(AX_\sigma^*)$ is selfadjoint and has the same zero space as $t^l(X_\sigma^*)$. But also,

$$AX_\sigma^*=X_\sigma A=J(-X_\sigma^*D)=J(-DX_\sigma),$$

where $J:\mathcal{U}_q\to\mathcal{U}_q$ is the involutive algebra isomorphism generated by

$$J(A)=D,\quad J(D)=A,\quad J(B)=C,\quad J(C)=B$$

(well defined in view of (3.2)). Also observe from (3.14) that

$$t^l_{m,n}(J(X))=t^l_{-m,-n}(X),\qquad X\in\mathcal{U}_q.$$

LEMMA 4.4. *$\sum_{m=-l}^{l}a_m\,e^l_m$ is eigenvector of $t^l(-DX_\sigma)$ with eigenvalue λ if and only if $\sum_{m=-l}^{l}a_m\,e^l_{-m}$ is an eigenvector of $t^l(AX_\sigma^*)$ with eigenvalue λ.*

Since $X_\sigma^*D=DX_\sigma$, we see also that $t^l(X_\sigma)\,v=0$ if and only if $t^l(X_\sigma^*)\,(t^l(D)\,v)=0$. In combination with Lemma 4.4 this yields the following.

LEMMA 4.5. *Let $l \in \frac{1}{2}\mathbb{Z}_+$ and $c_{-l}, c_{-l+1}, \dots, c_l \in \mathbb{C}$. Then*

$$t^l(X_\sigma)\left(\sum_{m=-l}^{l} q^{-\frac{1}{2}m} c_m e_m^l\right) = 0 \Longleftrightarrow t^l(X_\sigma^*)\left(\sum_{m=-l}^{l} q^{\frac{1}{2}m} c_m e_m^l\right) = 0$$
$$\Longleftrightarrow t^l(X_\sigma)\left(\sum_{m=-l}^{l} q^{-\frac{1}{2}m} c_{-m} e_m^l\right) = 0.$$

So, by Theorem 4.3 and Lemma 4.2 we get the following.

LEMMA 4.6. *$t^l(X_\sigma)$ and $t^l(X_\sigma^*)$ have zero-dimensional null space if $l \in \mathbb{Z}_+ + \frac{1}{2}$ and one-dimensional null space spanned by*

$$\sum_{m=-l}^{l} q^{-\frac{1}{2}m}\, c_m^{l,\sigma}\, e_m^l, \quad \textit{respectively,} \quad \sum_{m=-l}^{l} q^{\frac{1}{2}m}\, c_m^{l,\sigma}\, e_m^l$$

if $l \in \mathbb{Z}_+$. Here

$$(4.12)\quad c_m^{l,\sigma} := \frac{i^m\, q^{-(l+\sigma)m}\, q^{m^2/2}}{(q^2;q^2)_{l+m}^{\frac{1}{2}}\,(q^2;q^2)_{l-m}^{\frac{1}{2}}}\; {}_3\phi_2\left[\begin{matrix} q^{-2l+2m}, q^{-2l}, -q^{-2l-2\sigma} \\ q^{-4l}, 0 \end{matrix}; q^2, q^2\right] = c_{-m}^{l,\sigma}.$$

Furthermore, the subspace of (σ,τ)-spherical elements in $\mathcal{A}_q^l$ is zero-dimensional if $l \in \mathbb{Z}_+ + \frac{1}{2}$ and one-dimensional if $l \in \mathbb{Z}_+$. For $l \in \mathbb{Z}_+$, the (σ,τ)-spherical elements are spanned by

$$(4.13)\qquad \sum_{n,m=-l}^{l} q^{(n-m)/2}\, c_m^{l,\sigma}\, \overline{c_n^{l,\tau}}\, t_{n,m}^l\,.$$

The symmetry $c_m^{l,\sigma} = c_{-m}^{l,\sigma}$ in (4.12) follows from Lemma 4.5, but this symmetry can also be shown for the ${}_3\phi_2$ in (4.12) by iteration of [7, (3.2.3)].

We already found that the (σ,τ)-spherical elements in $\mathcal{A}_q^1$ were spanned by the element given by (4.7). Since, by Proposition 3.5, the l-fold tensor product of the representation t^1 will be a direct sum of irreducible representations equivalent to t^k, $k = 0, 1, \dots, l$; the polynomials of degree $\le l$ in $\rho_{\sigma,\tau}$ will certainly be (σ,τ)-spherical elements contained in $\oplus_{k=0}^{l}\mathcal{A}_q^l$. On the other hand, the algebra homomorphism $\pi_{\theta/2}^1 \colon \mathcal{A}_q \to \mathbb{C}$ (cf. (3.22)) sends $(\rho_{\sigma,\tau})^k$ to $(\cos\theta)^k$, so the monomials $(\rho_{\sigma,\tau})^k$ will be linearly independent in $\mathcal{A}_q$. So the element given by (4.13) must be a polynomial of degree l in $\rho_{\sigma,\tau}$. Thus we can state the following.

PROPOSITION 4.7. *The algebra of (σ,τ)-spherical elements in $\mathcal{A}_q$ is generated by $\rho_{\sigma,\tau}$ (given by (4.8)) and is, as a linear space, the direct sum of the (σ,τ)-spherical elements in $\mathcal{A}_q^l$ $(l = 0, 1, 2, \dots)$, which are spanned by*

$$(4.14)\qquad \sum_{n,m=-l}^{l} q^{(n-m)/2}\, c_m^{l,\sigma}\, \overline{c_n^{l,\tau}}\, t_{n,m}^l = P_l^{\sigma,\tau}(\rho_{\sigma,\tau}),$$

where $P_l^{\sigma,\tau}$ is a certain polynomial of degree l.

Apply $\pi_{-\theta/2}^1$ to both sides of (4.14). Then, by (3.23),

$$(4.15)\qquad \sum_{n=-l}^{l} c_n^{l,\sigma}\, \overline{c_n^{l,\tau}}\, e^{in\theta} = P_l^{\sigma,\tau}(\cos\theta).$$

Remark 4.8. Consider (4.15) with $\sigma = \tau$, together with (4.12). Compare it with (2.12), together with (2.14). Then we obtain

$$P_l^{\sigma,\sigma} = \frac{|c_l^{l,\sigma}|^2}{(q^{2l+2};q^2)_l}\, p_l(\,.\,; -q^{2\sigma+1}, -q^{-2\sigma+1}, q, q \mid q^2), \tag{4.16}$$

where p_l is an Askey–Wilson polynomial (2.2).

LEMMA 4.9. *We have*

$$h\big(P_{l'}^{\sigma,\tau}(\rho)^*\, P_l^{\sigma,\tau}(\rho)\big) = \delta_{l,l'}\, \frac{(1-q^2)q^{2l}}{1-q^{2(2l+1)}}\, P_l^{\sigma,\sigma}(\tfrac{1}{2}(q+q^{-1}))\, P_l^{\tau,\tau}(\tfrac{1}{2}(q+q^{-1})). \tag{4.17}$$

Proof. Apply (3.21) and (4.14). The case $l \neq l'$ is clear. For $l = l'$ we have

$$\begin{aligned} h\big(P_l^{\sigma,\tau}(\rho)^*\, P_l^{\sigma,\tau}(\rho)\big) &= \sum_{n,m=-l}^{l} q^{n-m}\, c_n^{l,\tau}\, \overline{c_m^{l,\sigma}}\, \overline{c_n^{l,\tau}}\, c_m^{l,\sigma}\, q^{2(l-n)}\, \frac{1-q^2}{1-q^{2(2l+1)}} \\ &= \frac{(1-q^2)q^{2l}}{1-q^{2(2l+1)}} \left(\sum_{m=-l}^{l} q^{-m}\, |c_m^{l,\sigma}|^2\right) \left(\sum_{n=-l}^{l} q^{-n}\, |c_n^{l,\tau}|^2\right) \\ &= \frac{(1-q^2)q^{2l}}{1-q^{2(2l+1)}}\, P_l^{\sigma,\sigma}(\tfrac{1}{2}(q+q^{-1}))\, P_l^{\tau,\tau}(\tfrac{1}{2}(q+q^{-1})). \qquad \square \end{aligned}$$

5. The action of the Casimir operator. Let $\lambda \in \mathbb{Z}_+$, let Ω be the Casimir element given by (3.7), let $\rho_{\sigma,\tau}$ be given by (4.8) and $P_l^{\sigma,\tau}$ by (4.15). We have

$$\begin{aligned} \langle A^\lambda \Omega, P_l^{\sigma,\tau}(\rho_{\sigma,\tau})\rangle &= \langle A^\lambda, \Omega.P_l^{\sigma,\tau}(\rho_{\sigma,\tau})\rangle \\ &= \left(\frac{q^{-l-\frac{1}{2}} - q^{l+\frac{1}{2}}}{q^{-1}-q}\right)^2 \langle A^\lambda, P_l^{\sigma,\tau}(\rho_{\sigma,\tau})\rangle \\ &= \left(\frac{q^{-l-\frac{1}{2}} - q^{l+\frac{1}{2}}}{q^{-1}-q}\right)^2 P_l^{\sigma,\tau}(\tfrac{1}{2}(q^\lambda + q^{-\lambda})), \end{aligned} \tag{5.1}$$

where the second identity follows from (3.20). Let X_σ be given by (4.1).

LEMMA 5.1. *We have*

$$\begin{aligned} q(q^{-1}-q)^2 A^\lambda \Omega \;\in\; & f(q^\lambda)\,(A^{\lambda+2} - A^\lambda) + f(q^{-\lambda})\,(A^{\lambda-2} - A^\lambda) \\ & + (1-q)^2 A^\lambda + \mathcal{U}_q X_\sigma + X_\tau \mathcal{U}_q\,, \end{aligned} \tag{5.2}$$

where

$$f(q^\lambda) := \frac{(1+q^{\sigma+\tau+1+\lambda})\,(1+q^{-\sigma-\tau+1+\lambda})\,(1-q^{\sigma-\tau+1+\lambda})\,(1-q^{-\sigma+\tau+1+\lambda})}{(1-q^{2\lambda})\,(1-q^{2\lambda+2})}\,. \tag{5.3}$$

Proof. If $Y, Z \in \mathcal{U}_q$, then $Y \sim Z$ will mean that

$$Y \;\in\; Z + \mathcal{U}_q X_\sigma + X_\tau \mathcal{U}_q\,.$$

First observe that

$$A^\lambda BC = q^\lambda (B - q^{-1}C) A^\lambda C + q^\lambda C A^\lambda (q^{-1}C - B) + q^{2\lambda} A^\lambda CB.$$

Substitute

$$A^\lambda CB = A^\lambda BC - A^\lambda(A^2 - D^2)/(q - q^{-1})$$

and

$$q^{\frac{1}{2}}B - q^{-\frac{1}{2}}C = -i\big(X_\tau + (q^{-\tau} - q^\tau)(A - D)/(q^{-1} - q)\big) \tag{5.4}$$

and similarly for X_σ. Then

$$\begin{aligned} iq^{\frac{1}{2}}(1 - q^{2\lambda})(q^{-1} - q)\,A^\lambda BC \sim{}& \big(q^{-1}(q^{-\tau} - q^\tau) - q^\lambda(q^{-\sigma} - q^\sigma)\big)\,CA^{\lambda+1} \\ &- \big(q(q^{-\tau} - q^\tau) - q^\lambda(q^{-\sigma} - q^\sigma)\big)\,CA^{\lambda-1} + iq^{2\lambda+\frac{1}{2}}\,(A^{\lambda+2} - A^{\lambda-2}). \end{aligned} \tag{5.5}$$

Observe that

$$CA^\mu = (C - qB)A^\mu + q^{1-\mu}\,A^\mu(B - q^{-1}C) + q^{-2\mu}\,CA^\mu.$$

Substitute again (5.4) and its analogue for X_σ; we obtain

$$(q^{-2\mu} - 1)(q^{-1} - q)\,CA^\mu \sim i\big(q^{\frac{1}{2}-\mu}(q^{-\sigma} - q^\sigma) - q^{\frac{1}{2}}(q^{-\tau} - q^\tau)\big)\,(A^{\mu+1} - A^{\mu-1}).$$

Substitute this last equivalence in (5.5). We obtain

$$\begin{aligned} &(1 - q^{2\lambda})(q^{-1} - q)^2\,A^\lambda BC \sim (q^{-2\lambda-2} - 1)^{-1} \\ &\quad\times \big(q^{-1}(q^{-\tau} - q^\tau) - q^\lambda(q^{-\sigma} - q^\sigma)\big)\,\big(q^{-1-\lambda}(q^{-\sigma} - q^\sigma) - (q^{-\tau} - q^\tau)\big)\,(A^{\lambda+2} - A^\lambda) \\ &\quad\quad - (q^{-2\lambda+2} - 1)^{-1}\,\big(q(q^{-\tau} - q^\tau) - q^\lambda(q^{-\sigma} - q^\sigma)\big) \\ &\quad\cdot \big(q^{1-\lambda}(q^{-\sigma} - q^\sigma) - (q^{-\tau} - q^\tau)\big)(A^\lambda - A^{\lambda-2}) + q^{2\lambda}(q^{-1} - q)\,(A^{\lambda+2} - A^{\lambda-2}). \end{aligned}$$

Now add $(1 - q^{2\lambda})\,A^\lambda(q^{-1}A^2 + qD^2 - 2)$ to both sides and multiply both sides with $q(1 - q^{2\lambda})^{-1}$. Then the left-hand side becomes $q(q^{-1} - q)^2\,A^\lambda\Omega$ and the right-hand side can be rewritten as

$$f(q^\lambda)\,(A^{\lambda+2} - A^\lambda) + f(q^{-\lambda})\,(A^{\lambda-2} - A^\lambda) + (1 - q)^2\,A^\lambda$$

with f given by (5.3). □

Substitute (5.2) into the left-hand side of (5.1). With the notation

$$R_l(q^\lambda) := P_l^{\sigma,\tau}(\tfrac{1}{2}(q^\lambda + q^{-\lambda}))$$

we obtain

(5.6)

$$f(q^\lambda)\,(R_l(q^{\lambda+2}) - R_l(q^\lambda)) + f(q^{-\lambda})\,(R_l(q^{\lambda-2}) - R_l(q^\lambda)) = -(1 - q^{-2l})(1 - q^{2l+2})\,R_l(q^\lambda).$$

Since this is an identity of rational functions in q^λ which is valid for infinitely many values of q^λ, it will remain valid if q^λ is arbitrarily complex, in particular if q^λ is replaced by $e^{i\theta}$. Then we recognize (5.6) as the second-order q-difference equation for Askey–Wilson polynomials; cf. (2.8). Hence

$$R_l(e^{i\theta}) = \text{const.}\,p_l(\cos\theta; -q^{\sigma+\tau+1}, -q^{-\sigma-\tau+1}, q^{\sigma-\tau+1}, q^{-\sigma+\tau+1} \mid q^2),$$

where p_l is an Askey–Wilson polynomial (2.2). We can compute the constant by comparing the coefficient of $e^{il\theta}$ at both sides (use (4.15)). The result generalizing (4.16) is the following.

THEOREM 5.2. *The polynomial $P_l^{\sigma,\tau}$ occurring in* (4.14) *and* (4.15) *equals*

$$P_l^{\sigma,\tau} = \frac{c_l^{l,\sigma}\,\overline{c_l^{l,\tau}}}{(q^{2l+2};q^2)_l}\, p_l(\,.\,; -q^{\sigma+\tau+1}, -q^{-\sigma-\tau+1}, q^{\sigma-\tau+1}, q^{-\sigma+\tau+1} \mid q^2). \tag{5.7}$$

THEOREM 5.3. *Let $dm(x) = dm_{a,b,c,d;q}(x)$ be the normalized orthogonality measure for the Askey–Wilson polynomials $p_n(x;a,b,c,d \mid q)$ as in* (2.7). *Let p be any polynomial. Then*

$$h(p(\rho_{\sigma,\tau})) = \int p(x)\, dm_{a,b,c,d;q^2}(x), \tag{5.8}$$

where $a = -q^{\sigma+\tau+1}$, $b = -q^{-\sigma-\tau+1}$, $c = q^{\sigma-\tau+1}$, $d = q^{-\sigma+\tau+1}$.

Proof. By (4.17) (for $l' = 0$) and (5.7) the theorem is therefore valid for $p = P_l^{\sigma,\tau}$ ($l \in \mathbb{Z}_+$). □

In the proof of Theorem 5.3 we only used the case $l' = 0$ of (4.17). Substitution of (5.7) and (5.8) into (4.17) for general l, l' should yield the full orthogonality relations for the Askey–Wilson polynomials of these particular parameters. We can indeed check that this is true. For $l = l'$ the left-hand side of (4.17) becomes

$$\frac{|c_l^{l,\sigma}|^2\,|c_l^{l,\tau}|^2}{(q^{2l+2};q^2)_l^2} \int p_l(x;a,b,c,d;q^2)^2\, dm_{a,b,c,d;q^2}(x),$$

where a, b, c, d are as in Theorem 5.3, while the right-hand side becomes

$$\frac{|c_l^{l,\sigma}|^2\,|c_l^{l,\tau}|^2\,(q^2;q^2)_l^2}{(q^{2l+2};q^2)_l^2\, q^{4l}}\,(-q^{2\sigma+2}, -q^{-2\sigma+2}, -q^{2\tau+2}, -q^{-2\tau+2}; q^2)_l\,.$$

These two expressions are equal because of (2.4), (2.5), and (2.6).

Remark 5.4. It follows from (4.15), (4.12), and (5.7) that

$$\begin{aligned} &p_n(\cos\theta; -q^{(\sigma+\tau+1)/2}, -q^{(-\sigma-\tau+1)/2}, q^{(\sigma-\tau+1)/2}, q^{(-\sigma+\tau+1)/2} \mid q) \\ &\quad = \sum_{k=-n}^{n} \frac{(q^{n+1};q)_n\,(q;q)_{2n}}{(q;q)_{n+k}(q;q)_{n-k}}\, q^{(n-k)(n-k+\sigma+\tau)/2} \\ &\qquad \times {}_3\phi_2\left[\begin{matrix} q^{-n+k}, q^{-n}, -q^{-n-\sigma} \\ q^{-2n}, 0 \end{matrix}; q, q\right] {}_3\phi_2\left[\begin{matrix} q^{-n+k}, q^{-n}, -q^{-n-\tau} \\ q^{-2n}, 0 \end{matrix}; q, q\right] e^{ik\theta}. \end{aligned} \tag{5.9}$$

This formula, obtained from the quantum group interpretation, cannot be found in the literature in the case of general σ, τ. For $\sigma = \tau$ we already gave an analytic proof of (5.9) in (2.12), (2.14). In a forthcoming paper [9] we will give an analytic proof for (5.9) in general and even for an extension of it with one more parameter. There it will turn out that the addition formula for classical ultraspherical polynomials (for Legendre polynomials in the case of (5.9)) is a limit case of our result. So it may be considered as an alternative to the Rahman–Verma [19] addition formula. In fact its derivation will be similar as for the Rahman–Verma formula.

6. Little and big q-Jacobi polynomials as limit cases of Askey–Wilson polynomials. Propositions 6.1 and 6.3 in this section are limit results for special functions, motivated by quantum group theory, but independent of quantum groups in formulation and proof. Before the author's paper [11] these limits have not been mentioned in literature, although R. Askey told me that he had been aware of them already some years ago.

Let X_σ ($\sigma \in \mathbb{R}$) be given by (4.1). Let $\mathcal{B}_\sigma := \{a \in \mathcal{A}_q \mid X_\sigma.a = 0\}$. By Lemma 3.2, $\mathcal{B}_\sigma$ is a subalgebra of $\mathcal{A}_q$; and, by (3.12) and (4.2), $\mathcal{B}_\sigma$ is moreover a $*$-subalgebra. It follows from (3.11) that $\Delta(\mathcal{B}_\sigma) \subset \mathcal{A}_q \otimes \mathcal{B}_\sigma$. Thus the quantum group $SU_q(2)$ corresponding to the Hopf $*$-algebra $\mathcal{A}_q$ acts on the quantum space corresponding to the $*$-algebra $\mathcal{B}_\sigma$. Thus it is natural to conjecture that this quantum action of $SU_q(2)$ coincides with its action on some quantum sphere as considered by Podleś [18]. According to Noumi and Mimachi [17, §5] this is indeed the case and they have made a precise identification between the two models.

Here we will restrict ourselves to the question of finding the elementary (σ, ∞)-spherical elements in $\mathcal{A}_q$. These can also be characterized as the elements of $\mathcal{B}_\sigma$ belonging to irreducible subspaces (with respect to $SU_q(2)$) and being invariant with respect to the diagonal quantum subgroup of $SU_q(2)$. We will obtain these (σ, ∞)-spherical elements as limit cases for $\tau \to \infty$ of the corresponding (σ, τ)-spherical elements.

It follows from Proposition 4.7, Theorem 5.2, and (2.3) that the (σ, ∞)-spherical elements in $\mathcal{A}_q^l$ ($l = 0, 1, 2, \ldots$) are spanned by

$$\lim_{\tau \to \infty} r_l(\rho_{\sigma,\tau}; q^{\tau-\sigma+1}, q^{-\tau+\sigma+1}, -q^{-\tau-\sigma+1}, -q^{\tau+\sigma+1} \mid q^2), \tag{6.1}$$

provided this limit exists and is nonzero. The limit can be obtained from the following limit transition from general Askey–Wilson polynomials to general big q-Jacobi polynomials.

PROPOSITION 6.1. *Let Askey–Wilson polynomials and big q-Jacobi polynomials be denoted by* (2.3) *and* (2.16), *respectively. Then*

$$\begin{aligned}\lim_{a \to 0} r_n\Bigg(\frac{q^{\frac{1}{2}}x}{2a(cd)^{\frac{1}{2}}}; q^{\alpha+\frac{1}{2}}a(d/c)^{\frac{1}{2}}, q^{\frac{1}{2}}a^{-1}(c/d)^{\frac{1}{2}},\\ -q^{\frac{1}{2}}a^{-1}(d/c)^{\frac{1}{2}}, -q^{\beta+\frac{1}{2}}a(c/d)^{\frac{1}{2}} \mid q\Bigg) = P_n^{(\alpha,\beta)}(x; c, d; q).\end{aligned} \tag{6.2}$$

Proof. The left-hand side of (6.2) can be written as

$$\sum_{k=0}^{n} \frac{(q^{-n}, q^{n+\alpha+\beta+1}; q)_k\, q^k}{(q^{\alpha+1}, -q^{\alpha+1}d/c, -q^{\alpha+\beta+1}a^2, q; q)_k} \prod_{j=0}^{k-1}\left(1 - \frac{q^{\alpha+1}x}{c}\, q^j + \frac{q^{2\alpha+1}a^2d}{c}\, q^{2j}\right). \quad \square$$

From (6.1), (6.2), and (4.7) we now obtain the following.

THEOREM 6.2. *The (σ, ∞)-spherical elements in $\mathcal{A}_q^l$ ($l = 0, 1, 2, \ldots$) are spanned by*

$$P_l^{(0,0)}(q^{-1}(1 - q^{2\sigma})\beta\gamma - iq^{\sigma-1}(\delta\beta + q\gamma\alpha); q^{2\sigma}, 1; q^2),$$

where $P_l^{(0,0)}$ is a big q-Jacobi polynomial.

The above theorem corresponds nicely with the interpretation of big q-Jacobi polynomials on quantum spheres by Noumi and Mimachi [16].

We can also try to get the (∞, ∞)-spherical elements in $\mathcal{A}_q^l$ by the limit

$$\lim_{\tau \to \infty} r_l(\rho_{-\tau,\tau}; q^{2\tau+1}, q^{-2\tau+1}, -q, -q \mid q^2). \tag{6.3}$$

For this we need the following.

PROPOSITION 6.3. *Let Askey–Wilson polynomials and little q-Jacobi polynomials be denoted by* (2.3) *and* (2.15), *respectively. Then*

$$\lim_{a\to 0} r_n\left(\frac{q^{\frac{1}{2}}x}{2a^2}; q^{\alpha+\frac{1}{2}}a^2, q^{\frac{1}{2}}a^{-2}, -q^{\frac{1}{2}}, -q^{\beta+\frac{1}{2}} \mid q\right) = \frac{(q^{\beta+1};q)_n}{(q^{-n-\alpha};q)_n}\, p_n(x; q^\beta, q^\alpha; q). \tag{6.4}$$

Proof. Put $d := a^2$ in the proof of Proposition 6.1. Then we obtain for the limit in (6.4)

$${}_3\phi_2\left[\begin{matrix} q^{-n}, q^{n+\alpha+\beta+1}, q^{\alpha+1}x \\ q^{\alpha+1}, 0 \end{matrix}; q, q\right].$$

Now the proposition follows from [7, (III.7)] and (2.15). □

From (6.3), (6.4), and (4.7) we now obtain the following.

THEOREM 6.4. *The* (∞,∞)*-spherical elements in* $\mathcal{A}_q^l$ $(l = 0, 1, 2, \ldots)$ *are spanned by*

$$p_l(-q^{-1}\beta\gamma; 1, 1; q^2),$$

where p_l *is a little q-Jacobi polynomial.*

The (∞,∞)-spherical elements in $\mathcal{A}_q^l$ coincide with the bi-invariant elements in $\mathcal{A}_q^l$ with respect to the diagonal quantum subgroup of $SU_q(2)$. These last ones are well known; cf. for instance [10], where we find the same explicit expression as in Theorem 6.3.

Remark 6.5. Askey–Wilson polynomials, with q fixed but with dilation of the argument admitted, form a five-parameter family of orthogonal polynomials. When these parameters are chosen as the α, β, a, c, d in the left-hand side of (6.2) then, for each choice of α, β, we obtain a three-parameter family of orthogonal polynomials which contain on the one hand the continuous q-Jacobi polynomials in Rahman's notation $P_n^{(\alpha,\beta)}(x;q)$ (cf. [5, (4.17)]) and on the other hand big and little q-Jacobi polynomials as limit cases.

Remark 6.6. When we compare Proposition 6.1 with Proposition 2.2 we see that the orthogonal polynomials in x after the limit sign in the left-hand side of (6.2) will have continuous mass on the interval $[-2a(cd/q)^{\frac{1}{2}}, 2a(cd/q)^{\frac{1}{2}}]$ and discrete mass points on the two sets

$$\{cq^k + a^2dq^{-k-1} \mid k \in \mathbb{Z}_+,\ q^k > a(qc/d)^{-\frac{1}{2}}\}$$

and

$$\{-dq^k - a^2cq^{-k-1} \mid k \in \mathbb{Z}_+,\ q^k > a(qd/c)^{-\frac{1}{2}}\}.$$

Clearly, when $a \to 0$, the continuous mass interval shrinks to $\{0\}$, while the discrete mass points tend two the two infinite sets $\{cq^k \mid k \in \mathbb{Z}_+\}$ and $\{-dq^k \mid k \in \mathbb{Z}_+\}$, just the location for the mass points of the big q-Jacobi polynomials. A similar remark can be made about the limit transition in Proposition 6.3.

REFERENCES

[1] G. E. ANDREWS AND R. ASKEY, *Enumeration of partitions: the role of Eulerian series and q-orthogonal polynomials*, in Higher Combinatorics, M. Aigner, ed., Reidel, Boston, MA, 1977, pp. 3–26.

[2] ———, *Classical orthogonal polynomials*, in Polynômes Orthogonaux et Applications, C. Brezinski, A. Draux, A. P. Magnus, P. Maroni, and A. Ronveaux, eds., Lecture Notes in Math., 1171, Springer, New York, 1985, pp. 36–62.

[3] R. Askey and M. E. H. Ismail, *A generalization of ultraspherical polynomials*, in Studies in Pure Mathematics, P. Erdös, ed., Birkhäuser, Basel, Switzerland, 1983, pp. 55-78.

[4] R. Askey and J. Wilson, *A set of orthogonal polynomials that generalize the Racah coefficients or 6-j symbols*, SIAM J. Math. Anal., 10 (1979), pp. 1008–1016.

[5] ———, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 54 (1985), no. 319.

[6] G. Gasper and M. Rahman, *A nonterminating q-Clausen formula and some related product formulas*, SIAM J. Math. Anal., 20 (1989), pp. 1270–1282.

[7] ———, *Basic hypergeometric series*, Encyclopedia of Mathematics and Its Applications, 35, Cambridge University Press, London, 1990.

[8] H. T. Koelink, *The addition formula for continuous q-Legendre polynomials and associated spherical elements on the $SU(2)$ quantum group related to Askey–Wilson polynomials*, SIAM J. Math. Anal., submitted.

[9] T. H. Koornwinder, *A second addition formula for continuous q-ultraspherical polynomials motivated by quantum groups*, in preparation.

[10] ———, *Representations of the twisted $SU(2)$ quantum group and some q-hypergeometric orthogonal polynomials*, Nederl. Akad. Wetensch. Proc. Ser. A, 92 (1989), pp. 97–117.

[11] ———, *Orthogonal polynomials in connection with quantum groups*, in Orthogonal Polynomials: Theory and Practice, P. Nevai, ed., NATO-ASI Series C, 294, Kluwer, Norwell, MA, 1990, pp. 257–292.

[12] T. Masuda, K. Mimachi, Y. Nakagami, M. Noumi, Y. Saburi, and K. Ueno, *Unitary representations of the quantum group $SU_q(1,1)$: structure of the dual space of $\mathcal{U}_q(sl(2))$*, Lett. Math. Phys., 19 (1990), pp. 187–194.

[13] T. Masuda, K. Mimachi, Y. Nakagami, M. Noumi, and K. Ueno, *Representations of the quantum group $SU_q(2)$ and the little q-Jacobi polynomials*, J. Funct. Anal., 99 (1991), pp. 127–151.

[14] ———, *Representations of quantum groups and a q-analogue of orthogonal polynomials*, C. R. Acad. Sci. Paris Sér. I Math., 307 (1988), pp. 559–564.

[15] M. Noumi and K. Mimachi, *Askey–Wilson polynomials and the quantum group $SU_q(2)$*, Proc. Japan Acad. Ser. A Math. Sci., 66 (1990), pp. 146–149.

[16] ———, *Quantum 2-spheres and big q-Jacobi polynomials*, Comm. Math. Phys., 128 (1990), pp. 521–531.

[17] ———, *Askey–Wilson polynomials as spherical functions on $SU_q(2)$*, in Quantum Groups, P. Kulish, ed., Lecture Notes in Math., 1510, Springer, New York, 1992, pp. 98–103.

[18] P. Podleś, *Quantum spheres*, Lett. Math. Phys., 14 (1987), pp. 193–202.

[19] M. Rahman and A. Verma, *Product and addition formula for the continuous q-ultraspherical polynomials*, SIAM J. Math. Anal., 17 (1986), pp. 1461–1474.

[20] D. Stanton, *Orthogonal polynomials and Chevalley groups*, in Special Functions: Group Theoretic Aspects and Applications, R. A. Askey, T. H. Koornwinder, and W. Schempp, eds., Reidel, Boston, MA, 1984, pp. 87–128.

[21] L. L. Vaksman and Ya. S. Soĭbel'man, *Algebra of functions on the quantum group $SU(2)$*, Functional Anal. Appl., 22 (1988), pp. 170–181.

SIAM J. MATH. ANAL.
Vol. 24, No. 3, pp. 814–823, May 1993

SOME INEQUALITIES FOR THE FIRST POSITIVE ZEROS OF BESSEL FUNCTIONS*

LEE LORCH†

Abstract. For the first positive zero $j = j_{\nu 1}$ of the Bessel function $J_\nu(x)$, it is shown for $-1 < \nu < \infty$, that (i) $j^2 - (\nu+1)(\nu+5)$ increases from zero to $+\infty$ so that $j^2 > (\nu+1)(\nu+5)$, and that

$$\text{(ii)}\qquad j^2 > \frac{24(\nu+1)^2}{1-2\nu+\sqrt{(2\nu+3)(2\nu+11)}} - 2(\nu^2-1).$$

These inequalities find use in the work of Ashbaugh and Benguria on bounding ratios of eigenvalues. In addition, some associated monotonicities and inequalities are established and some conjectures of Ismail and Muldoon on lower bounds for j are verified.

Key words. Bessel functions, zeros, inequalities, monotonicity

AMS subject classification. primary 33A40

1. Introduction and statement of principal results.

In the course of their work on bounding ratios of eigenvalues, including proofs of long-standing Payne–Pólya–Weinberger conjectures, Ashbaugh and Benguria [1], [2], [3], encountered the need for some inequalities for $j = j_{\nu 1}$, the first positive zero of the Bessel function $J_\nu(x)$.

They require the inequalities

$$(1)\qquad j^2 > (\nu+1)(\nu+5), \qquad -1 < \nu < \infty,$$

and

$$(2)\qquad j^2 > \frac{24(\nu+1)^2}{1-2\nu+\sqrt{(2\nu+3)(2\nu+11)}} - 2(\nu^2-1), \qquad -1 < \nu < \infty,$$

for the special values $\nu = \frac{1}{2}n - 1$, $n = 1, 2, \ldots$, where n is the dimension of the space in which the domains they study are embedded. To fill this need, proofs of (1) and (2) are among the results provided here.

A straightforward calculation shows that (2) provides a sharper (larger) lower bound for $j_{\nu 1}$ than does (1), i.e., that the right member of (1) is smaller than the right member of (2). Hence, it would suffice to prove only (2).

The inequality (2) can be rewritten readily as

$$1 + \frac{6(\nu+1)}{j^2 + 2(\nu^2-1)} < \frac{2\nu+5+\sqrt{(2\nu+3)(2\nu+11)}}{4(\nu+1)}, \qquad \nu > -1.$$

With $\nu = \frac{1}{2}n - 1$, this becomes

$$(2')\qquad 1 + \frac{6n}{2j^2 + n(n-4)} < \frac{n+3+\sqrt{(n+1)(n+9)}}{2n}, \qquad n > 0.$$

* Received by the editors March 24, 1992; accepted for publication (in revised form) August 17, 1992. The Natural Sciences and Engineering Research Council of Canada provided partial support for this work.

† Department of Mathematics, York University, North York, Ontario, Canada M3J 1P3.

In this form, the inequality establishes that Chiti's upper bound for $(j_{n/2}/j_{n/2-1})^2$ is more precise (i.e., smaller) than the upper bound for that ratio found by Brands and also jointly by Hile and Protter. This is discussed and references are provided in [2]. Here $j_\nu = j_{\nu 1}$.

However, a separate analysis of (1) will be included, inferring its validity from the more informative result that

$$f(\nu) = j^2 - (\nu+1)(\nu+5) \tag{3}$$

increases steadily from zero (since $f(-1^+) = 0$) to $+\infty$, $-1 < \nu < \infty$.

This monotonicity is obviously equivalent to the inequality

$$j\frac{dj}{d\nu} > \nu + 3, \qquad -1 < \nu < \infty, \tag{4}$$

which may have further interest because of its simplicity; it is in this formulation that the monotonicity of $f(\nu)$, and hence also the truth of the inequality (1), will be demonstrated.

The inequality (1) was formulated by Elbert [4], who proved it for $-1 < \nu < 0$, indicated a proof for $0 \le \nu < 1$, and conjectured its validity for $0 \le \nu < \infty$.

2. Collateral results. The proof of (4), beginning with inequalities found by Ismail and Muldoon [8], invites attention to various conjectures in [8]. These will be verified below (§6).

3. Proof of inequality (4). This proof involves a consideration of three subintervals (the first two overlap) of $-1 < \nu < \infty$:

$$\text{(A)} \quad -1 < \nu \le 1.4, \quad \text{(B)} \quad 0 \le \nu < 2, \quad \text{and} \quad \text{(C)} \quad 2 \le \nu < \infty.$$

In (A), inequalities due to Ismail and Muldoon [8] will yield the desired conclusion. In (B), the result follows from Elbert's concavity theorem $d^2j/d^2\nu < 0$, $\nu \ge 0$ [5] and the Elbert–Laforgia inequality $dj/d\nu > 1, \nu \ge 0$ [6]. In (C) the Elbert–Laforgia inequality will play the essential role.

Proof for (A). From [8, (6.2)], (4) will hold for $-1 < \nu \le 1.4$ if

$$2\left[1 + \frac{4(\nu+1)^2}{j^2}\right] \ge \nu + 3, \qquad -1 < \nu \le 1.4,$$

i.e., if $8(\nu+1) \ge j^2$, $-1 < \nu \le 1.4$.

This in turn will be the case, according to [8, (6.10)], if

$$8(\nu+1) \ge \frac{2(\nu+1)(\nu+5)(5\nu+11)}{7\nu+19}, \qquad -1 < \nu \le 1.4,$$

i.e., if $0 \ge 5\nu^2 + 8\nu - 21 = (\nu+3)(5\nu-7)$, which holds for this ν-interval.

Proof for (B). $0 \le \nu < 2$. Here, and below in the proof of (4′), we rely on Elbert's concavity theorem. Accordingly,

$$\left.\frac{dj}{d\nu}\right|_{\nu=\rho} > \left.\frac{dj}{d\nu}\right|_{\nu=\mu} > \frac{j_{\mu+h} - j_\mu}{h}, \quad \mu > \rho \ge 0, \quad h > 0. \tag{5}$$

Putting $\mu = 2$ and $h = \frac{1}{2}$, we have

$$\frac{dj}{d\nu} > 1.2556738, \qquad 0 \le \nu < 2,$$

so that

$$j\frac{dj}{d\nu}-\nu > 1.2556738j-\nu, \qquad 0 \le \nu < 2.$$

The right side is an increasing function of ν, since $dj/d\nu > 1$, $\nu \ge 0$. This lower bound exceeds 3 already when $\nu = 0$.

Proof for (C). Similarly, (4) holds if $j-\nu \ge 3$. This is the case for $\nu = 2$, since $j_{2,1} > 5$, and hence also for all $\nu > 2$, since $j_\nu - \nu$ is an increasing function in $0 \le \nu < \infty$.

This concludes the proof of (4), hence also of (1) and of the monotonicity of (3) for all $-1 < \nu < \infty$.

4. Remarks on the proof. (1) Martin Muldoon has pointed out in conversation that (1) by itself provides a lower bound smaller than that established in [8, (6.8)] for $-1 < \nu < 7$. If the proof of (1) were the only objective, Muldoon's observation would permit deleting the proofs of (A) and (B) above, retaining only the brief proof of case (C).

(2) Inequality (4), stated and proved for $j = j_{\nu 1}$, holds all the more for $j = j_{\nu k}$, $k = 2,3,\dots$, and, more generally, for $j = j_{\nu\kappa}$, $\kappa \ge 1$, defined in [6]. In particular, $j_{\nu\kappa} = j_{\nu k}$, $\kappa = k = 1,2,\dots$, and $j_{\nu\kappa} = y_{\nu k}$ when $\kappa = k-\frac{1}{2}$, $k = 1,2,3,\dots$, where $y_{\nu k}$ is the kth positive zero of $Y_\nu(x)$, the Bessel function of second kind. This extension follows from (4) on noting that, for fixed ν, (i) $j_{\nu\kappa}$ increases with κ (cf. [6]) and (ii) $dj_{\nu\kappa}/d\nu$ also increases with κ from the lemma in [11].

5. Proof of inequality (2). This will be proved in two overlapping installments: (A) $-1 < \nu \le 15$ and (B) $1.34 \le \nu < \infty$. The proof of (B), as noted below, will establish results (1′) and (3′), strengthening (1) and (3) in the interval $1.34 \le \nu < \infty$ and also (4′), which strengthens (4) in the interval $.05 < \nu < \infty$.

(A) For these ν, the already established Ismail–Muldoon lower bound [8, (6.9)] for j^2 exceeds the right member of (2), i.e., after dividing each by 2,

$$\begin{aligned}2^{2/3}[(\nu+2)(\nu+3)]^{1/3} &> \frac{12(\nu+1)}{1-2\nu+\sqrt{(2\nu+3)(2\nu+11)}}-\nu+1\\ &= \frac{3}{8}\sqrt{(2\nu+3)(2\nu+11)}-\frac{1}{4}\nu+\frac{5}{8}.\end{aligned}$$

This inequality holds for $-1 < \nu \le 15.28966\dots$.

To verify this, we multiply both sides of the inequality by 8 and then cube each side. This yields an equivalent inequality which, simplified, becomes

$$56\nu^3+740\nu^2+2098\nu+1927 > 9(2\nu+3)^{1/2}(2\nu+11)^{1/2}(4\nu^2+16\nu+31).$$

Squaring each side of this inequality, then collecting all terms and finally dividing by 256, leads to another inequality, again equivalent to (2), namely,

$$-8\nu^6+20\nu^5+1118\nu^4+5915\nu^3+13067\nu^2+12713\nu+4471 > 0,$$

which may be written as

$$(\nu+1)^2(-8\nu^4+36\nu^3+1054\nu^2+3771\nu+4471) > 0,$$

and, since $\nu > -1$, finally as

$$-8\nu^4+36\nu^3+1054\nu^2+3771\nu+4471 > 0.$$

This quartic has exactly one positive root, ν_0, which could be written in "exact form." Numerically, $\nu_0 = 15.28966\ldots$.

Thus, (2) has been proved for $-1 < \nu \leq \nu_0 = 15.28966\ldots$.

Remark. This shows that the Ismail–Muldoon lower bound in [8, (6.9)] for j^2 improves on the bound provided by (2) when $-1 < \nu < \nu_0$. The position is reversed for $\nu > \nu_0$, as is evident from the foregoing discussion.

(B) The proof of (2) for $\nu > \nu_0$ will in fact cover the larger interval $\nu \geq 1.34$, since doing so presents an opportunity to establish results yielding additional information. This will permit improving (1), (3), and (4) for appropriate infinite ν-intervals, although not for the entire interval $-1 < \nu < \infty$.

The inequality (2) can be rewritten in the form

$$(2'') \qquad \frac{j^2}{2(\nu+1)} > \frac{3}{8}\sqrt{(2\nu+3)(2\nu+11)} - \frac{1}{4}\nu + \frac{5}{8}.$$

Now, $(2\nu+3)(2\nu+11) < (2\nu+7)^2$ so that the right side of (2″) is less than $\frac{1}{2}(\nu + 13/2)$. Thus, (2″), and hence (2) would follow for those $\nu > -1$ for which

$$(1') \qquad j^2 > (\nu+1)\left(\nu + \frac{13}{2}\right),$$

an inequality stronger than (1) for those $\nu > -1$ for which it is valid.

Paralleling the discussion of (1), it will be shown that (1′) *holds for* $\nu \geq 1.34$ (and hence also (2) and (2″)). It is reversed for $\nu = 1.33$ [9, p. 195].

The inequality (1′) follows, since $f_1(1.34) > 0$ [9, p. 195], from the property that

$$(3') \qquad f_1(\nu) = j^2 - (\nu+1)\left(\nu + \frac{13}{2}\right)$$

is an increasing function of ν *for* $.05 \leq \nu < \infty$. However, $f_1(.04) > f_1(.05)$.

The monotonicity assertion concerning $f_1(\nu)$ is equivalent to the inequality

$$(4') \qquad j\frac{dj}{d\nu} - \nu > \frac{15}{4}, \qquad .05 \leq \nu < \infty,$$

an inequality which, by the way, reverses when $\nu = 0$.

The proof of (4′) is separated into four parts, the first of which alone suffices, together with (A), to complete the proof of (2) and (2″).

(i) $5 \leq \nu < \infty$. Since $dj/d\nu > 1$, $\nu \geq 0$, we have

$$j\frac{dj}{d\nu} - \nu > j - \nu,$$

an increasing function of $\nu \geq 0$. But $j - \nu > 15/4$ when $\nu = 5$, so that this part of the proof of (4′) is complete—and with it the proof of (2) and (2″) is also complete.

(ii) $3/2 \leq \nu < 5$. Putting $\mu = 5$ and $h = \frac{1}{2}$ in (5), we have

$$j\frac{dj}{d\nu} - \nu > 1.16865654j - \nu, \qquad 0 \leq \nu < 5,$$

an increasing function of ν. When $\nu = 3/2$, this lower bound exceeds 15/4, thus verifying (4′) also for $3/2 \leq \nu < 5$.

(iii) $.78 \leq \nu < 3/2$. Following the pattern of (ii),

$$\frac{dj}{d\nu} > \frac{j_{2,1} - j_{1.5,1}}{.5} > 1.28442564, \qquad \nu < \frac{3}{2},$$

and so

$$j\frac{dj}{d\nu} - \nu > 1.28442564j - \nu, \qquad \nu < \frac{3}{2},$$

an increasing function of ν which exceeds 15/4 [9, p. 195] when $\nu = .78$. This verifies (4′) also for $.78 \leq \nu < 3/2$.

(iv) $.05 \leq \nu < .78$. For this portion of the ν-interval, there is available a larger lower bound for $dj/d\nu$ than the one given by (5). From the mean-value theorem,

$$\frac{dj}{d\nu} = \left.\frac{dj}{d\nu}\right|_{\nu=0} + \nu\left.\frac{d^2j}{d\nu^2}\right|_{\nu=\mu}$$

for some $\mu, 0 < \mu < \nu$. Elbert and Laforgia [7] have shown that $d^3j/d\nu^3 > 0,\ \nu \geq 0$, so that $d^2j/d\nu^2$ is an increasing function of ν. Thus, we now have

$$\frac{dj}{d\nu} > \left.\frac{dj}{d\nu}\right|_{\nu=0} + \nu\left.\frac{d^2j}{d\nu^2}\right|_{\nu=0}, \qquad \nu > 0.$$

Using the numerical values [10, (1.1)] now yields the lower bound to be used here, i.e.,

$$\frac{dj}{d\nu} > 1.54288974 - .175493593\nu, \nu > 0. \tag{6}$$

The argument is now subdivided into several steps.

(a) $.34 \leq \nu < .78$. Here $dj/d\nu > 1.40600473$ so that

$$j\frac{dj}{d\nu} - \nu > 1.40600473j - \nu, \qquad 0 \leq \nu < .78,$$

an increasing function of ν which exceeds 15/4 already when $\nu = .34$ [9, p. 195].

(b) $.15 \leq \nu < .34$. Similarly, for this interval,

$$j\frac{dj}{d\nu} - \nu > 1.48322192j - \nu,$$

an increasing function of ν which exceeds 15/4 already when $\nu = .15$ [9, p. 195].

(c) The procedure is now clear. It can be applied successively to the intervals $.08 \leq \nu < .15, .06 \leq \nu < .08, .05 \leq \nu < .06$, using, as above, the values for $j_{\nu 1}$ recorded in [9, p. 195].

Remark. As noted in §1, (2) implies (1).

6. On some Ismail–Muldoon bounds. (1) Ismail and Muldoon observed numerically [8, p. 201] that their lower bound for $j_{\nu 1}^4$ in [8, (6.27)] is better (larger) than the lower bound given by [8, (6.24)], when $\nu > 0$.

Their implied conjecture thus suggests that the following inequality holds between these respective lower bounds for $j_{\nu 1}^4$:

$$\begin{aligned} &j_{01}^4 + \frac{56}{3}\nu^3 + 8(j_{01} + 4)\nu^2 + 8(j_{01}^2 + 10)\nu - 128 \ln\left(1 + \frac{\nu}{3}\right) \\ &\quad > j_{01}^4 + \frac{40}{3}\nu^3 + 8(j_{01} + 4)\nu^2 + 8(j_{01}^2 + 4)\nu, \qquad \nu > 0, \end{aligned}$$

i.e.,

$$3g(\nu) := \nu^3 + 9\nu - 24 \ln\left(1 + \frac{\nu}{3}\right) > 0, \qquad \nu > 0.$$

Now,

$$g'(\nu) = \frac{(\nu+3)(\nu^2+3) - 8}{\nu+3} > 0, \qquad \nu > -1,$$

while $g(0) = 0$, and so $g(\nu) > 0,\ \nu > 0$.

This verifies the Ismail–Muldoon conjecture and shows also that the bound in [8, (6.27)] becomes increasingly better than the one provided by [8, (6.24)], $0 < \nu < \infty$.

(2) Ismail and Muldoon observed numerically [8, p. 200] that the lower bound they provide in [8, (6.26)] is larger than the one they give in [8, (6.23)], thereby implying the conjecture that this ordering would hold for all $\nu > -1$. This is the case. These lower bounds for $j_{\nu 1}^4$ are, respectively,

$$\frac{96}{5}\left[1 + \nu\left\{1 + \frac{2}{3}(\nu+1)\right\}^{3/2}\right] + 16\nu^3 + 32\nu^2 + 80\nu + 64 - 128 \ln\left\{\frac{1}{2}(\nu+3)\right\}$$

and

$$\frac{96}{5}\nu\left[1 + \frac{2}{3}(\nu+1)\right]^{3/2} + \frac{96}{5} + \frac{32}{3}(\nu+1)^3.$$

Showing that the first of these exceeds the second for all $\nu > -1$ is equivalent to verifying that

$$F(\nu) := \nu^3 + 9\nu + 10 - 24 \ln\left\{\frac{1}{2}(\nu+3)\right\} > 0, \qquad \nu > -1.$$

Now, $F(-1) = 0$, while $F'(\nu) > 0,\ \nu > -1$, so that $F(\nu) > 0,\ \nu > -1$, thereby establishing the correctness of this conjecture.

(3) Another Ismail–Muldoon conjecture [8, line following (6.18), p. 198] (which they supported numerically) that can be confirmed readily is that Schafheitlin's upper bound for $j_{\nu 1}$, given by [8, (6.16)] is sharper (i.e., less) than the one provided by [8, (6.10)] for all $\nu > -1$. That is, for $\nu > -1$,

$$\frac{2(\nu+1)(\nu+5)}{1 + [1 - \frac{3}{4}(\nu+1)(\nu+5)/\{(\nu+2)(\nu+4)\}]^{1/2}} < \frac{2(\nu+1)(\nu+5)(5\nu+11)}{7\nu+19},$$

these being the respective upper bounds for $j_{\nu 1}^2$ recorded in [8, (6.16)] and [8, (6.10)].

By cross-multiplying and simplifying, this becomes

$$2(\nu+4) < (5\nu+11)\left[1 - \frac{3(\nu+1)(\nu+5)}{4(\nu+2)(\nu+4)}\right]^{1/2}.$$

Squaring both sides and simplifying leads to the equivalent (and obvious) inequality $9(\nu+1)^4 > 0, \nu > -1$.

(4) Another essentially correct Ismail–Muldoon conjecture [8, p. 198] states that Schafheitlin's upper bound for $j_{\nu 1}^2, A/(1+B)$, recorded in [8, (6.16)], is sharper (smaller) than the upper bound,

$$\frac{8(7\nu+19)(\nu+1)(\nu+2)(\nu+3)(\nu+6)}{42\nu^3 + 362\nu^2 + 1026\nu + 946}$$

provided by [8, (6.11)] when $\nu > -0.54$. Here

$$A = 2(\nu+1)(\nu+5),$$
$$B = [1 - \tfrac{3}{4}(\nu+1)(\nu+5)/\{(\nu+2)(\nu+4)\}]^{1/2}.$$

More precisely, it will be shown that there exists a unique $\rho = -0.534\ldots$ such that the difference of these bounds,

$$G(\nu) = \frac{A}{1+B} - \frac{8(7\nu+19)(\nu+1)(\nu+2)(\nu+3)(\nu+6)}{42\nu^3+362\nu^2+1026\nu+946}, \tag{7}$$

is negative for $\nu > \rho$. On the other hand, $G(\nu) > 0$ for $-1 < \nu < \rho$, so that [8, (6.11)] is better than [8, (6.16)] for $-1 < \nu < \rho$. When $\nu = \rho$, $G(\nu) = 0$.

Now,

$$\begin{aligned} G(\nu)H(\nu) &= (\nu+5)(21\nu^3+181\nu^2+513\nu+473) \\ &\quad - \left(2 + \sqrt{\frac{\nu^2+6\nu+17}{(\nu+2)(\nu+4)}}\right)(7\nu+19)(\nu+2)(\nu+3)(\nu+6), \end{aligned}$$

where

$$H(\nu) = \left(2 + \sqrt{\frac{\nu^2+6\nu+17}{(\nu+2)(\nu+4)}}\right)\frac{21\nu^3+181\nu^2+513\nu+473}{4(\nu+1)}.$$

Clearly, $H(\nu) > 0$, $-1 < \nu < \infty$, so that $G(\nu)$ and $G(\nu)H(\nu)$ have the same signs in $-1 < \nu < \infty$. Now,

$$\begin{aligned} (\nu+4)^{1/2}G(\nu)H(\nu) &= (\nu+4)^{1/2}(7\nu^4+94\nu^3+496\nu^2+1766\nu+996) \\ &\quad - (7\nu+19)(\nu+3)(\nu+6)(\nu+2)^{1/2}(\nu^2+6\nu+17)^{1/2} \\ &:= G_1(\nu) - G_2(\nu), \end{aligned}$$

whence $G(\nu)$ has the same signs in $-1 < \nu < \infty$ as does

$$G_1^2(\nu) - G_2^2(\nu) = -(\nu+1)^4(\nu+5)(28\nu^3+191\nu^2+371\nu+148). \tag{8}$$

Thus, the signs of $G(\nu)$ in $-1 < \nu < \infty$ are opposite (except when zero) to those of

$$\psi(\nu) = 28\nu^3+191\nu^2+371\nu+148. \tag{9}$$

Following the classical method of solving cubic equations, we put $\nu = x-(191/84)$ so that

$$\psi(\nu) = \frac{8}{84^3}q(x) = \frac{8}{84^3}[2073214x^3 - 4689594x - 2759281]. \tag{10}$$

The cubic $q(x)$ has precisely one real root ξ; this can be expressed in exact form, namely,

$$\begin{aligned} \xi = &\left\{\frac{2759281}{4146428} + \sqrt{\left(\frac{2759281}{4146428}\right)^2 - \left(\frac{781599}{1036607}\right)^3}\right\}^{1/3} \\ &+ \left\{\frac{2759281}{4146428} - \sqrt{\left(\frac{2759281}{4146428}\right)^2 - \left(\frac{781599}{1036607}\right)^3}\right\}^{1/3}. \end{aligned}$$

Thus, $\psi(\nu)$ has exactly one real root ρ, where

$$\rho = \xi - \frac{191}{84} = -0.534\ldots. \tag{11}$$

It is easy to see that $\psi(\nu) < 0$, and hence $G(\nu) > 0$, $-1 < \nu < \rho$, while $\psi(\nu) > 0$ so that $G(\nu) < 0$ for $\rho < \nu < \infty$ and $G(\rho) = 0$.

Remark. The inequalities (bounds) on $j_{\nu 1}$ and $jdj/d\nu$ borrowed from [8] to establish various inequalities above are not the sharpest provided in [8]. However, the sharper (and more complicated) bounds found in [8] do not appear to yield stronger inequalities of the type discussed herein.

7. Miscellaneous comments. Inequalities (1) and (4) are each "best possible" in a certain sense. Increasing the factor $\nu + 5$ by any fixed positive quantity reverses (1) in a neighbourhood of $\nu > -1$. This is the case as well if (1) is rewritten as $j^2 > \nu^2 + 6\nu + 5$. Increasing the lower bound by a positive constant would again reverse the inequality in a neighbourhood of $\nu = -1$. The same is true of (4) if the term $\nu + 3$ is increased by a fixed amount. For example, given $a > 0$,

$$j^2 < (\nu + 1)(\nu + 5 + a), \qquad -1 < \nu \leq a - 1 \tag{12}$$

and

$$j\frac{dj}{d\nu} < \nu + 3 + \frac{1}{2}a, \qquad -1 < \nu \leq -1 + \varepsilon. \tag{13}$$

Similarly, the monotonicity of (3) is "best possible" in that

$$f_a(\nu) := j^2 - (\nu + 1)(\nu + 5 + a), \qquad a > 0,$$

while ultimately increasing, decreases in $-1 < \nu \leq -1 + \varepsilon$.

To prove (12), we observe that, for $-1 < \nu \leq a - 1$,

$$(\nu + 1)(\nu + 5 + a) > 2(\nu + 1)(\nu + 3) > j^2.$$

The second inequality is a standard upper bound for j^2, recorded, e.g., in [8, (6.8)]. As to (13), it should be noted that $f_a(-1^+) = 0$ and that $f_a(\nu) < 0$, $-1 < \nu \leq a - 1$, from (12). This implies the comments about $f_a(\nu)$, and hence also (13), except for the ultimate increasing character of $f_a(\nu)$. This latter follows from the inequality $dj/d\nu > 1$ since $j - \nu$ increases and (as is clear from the well-known asymptotics of $j_{\nu 1}$ [9, p. 153, §5.3]) becomes infinite as $\nu \to \infty$ so that

$$j\frac{dj}{d\nu} > j > \nu + 3 + \frac{1}{2}a$$

for ν sufficiently large.

Also, for any $A > 0$, the inequality

$$j^2 > (\nu + 1)(\nu + A) \tag{14}$$

holds for all sufficiently large ν, depending on A, again as a consequence of the monotonicity of

$$j^2 - (\nu + 1)(\nu + A).$$

To establish this monotonicity, it suffices to note that

$$j\frac{dj}{d\nu} - \nu - \frac{1}{2}(A+1) > j - \nu - \frac{1}{2}(A+1)$$

becomes and remains positive since $j - \nu$ increases to $+\infty$ as $\nu \to \infty$.

Other inequalities discussed in previous sections are also "best possible" in the intervals in which they hold.

Another inequality might be useful for cases, such as those considered in [1], [2], [3], where $\nu \geq \frac{1}{2}$:

$$j^2 > (\nu+1)\left(\nu + \frac{2}{3}\pi^2 - \frac{1}{2}\right), \qquad \frac{1}{2} < \nu < \infty. \tag{15}$$

This becomes an equality when $\nu = \frac{1}{2}$.

The inequality is a consequence of the monotonicity of

$$\varphi(\nu) = j^2 - (\nu+1)\left(\nu + \frac{2}{3}\pi^2 - \frac{1}{2}\right), \qquad 0 \leq \nu < \infty,$$

since $\varphi(\frac{1}{2}) = 0$.

The proof that $\varphi(\nu)$ increases in $0 \leq \nu < \infty$ follows the same lines as for (4) and (4′). We have

$$\frac{1}{2}\varphi'(\nu) = j\frac{dj}{d\nu} - \nu - \frac{1}{3}\pi^2 - \frac{1}{4}.$$

(i) $4 \leq \nu < \infty$. Since $dj/d\nu > 1$, so that $\frac{1}{2}\varphi'(\nu) > j - \nu - \frac{1}{3}\pi^2 - \frac{1}{4}$, an increasing function of ν which is positive already when $\nu = 4$.

(ii) $1 \leq \nu < 4$. Here we use (5) with $\mu = 4, h = \frac{1}{2}$ so that $dj/d\nu > 1.18843804$. Thus,

$$\frac{1}{2}\varphi'(\nu) > 1.18843804j - \left(\frac{1}{3}\pi^2 + \frac{1}{4}\right), \qquad 1 \leq \nu < 4,$$

an increasing function of ν which is positive already when $\nu = 1$.

(iii) $0 \leq \nu < 1$. This remaining interval is subdivided into the subcases, $\frac{1}{2} \leq \nu < 1$, $\frac{1}{4} \leq \nu < \frac{1}{2}$, $0 \leq \nu < \frac{1}{4}$, which are treated successively in the same fashion as (ii).

Acknowledgment. Preprints of Ashbaugh and Benguria's work [1], [2], [3] were kindly shown to me by Martin Muldoon.

REFERENCES

[1(a)] M. S. Ashbaugh and R. D. Benguria, *Proof of the Payne–Pólya–Weinberger conjecture*, Bull. Amer. Math. Soc., 25 (1991), pp. 19–29.

[1(b)] ———, *Sharp bound for the ratio of the first two eigenvalues of Dirichlet Laplacians and extensions*, Ann. Math., 135 (1992), pp. 601–628.

[2] ———, *More bounds on eigenvalue ratios for Dirichlet Laplacians in N dimensions*, preprint, 1991.

[3] ———, *A second proof of the Payne–Pólya–Weinberger conjecture*, Comm. Math. Phys., 147 (1992), pp. 181–190.

[4] Á. Elbert, *Some inequalities concerning Bessel functions of the first kind*, Studia Sci. Math. Hungar., 6 (1971), pp. 277–283.

[5] Á. Elbert, *Concavity of the zeros of Bessel functions*, Studia Sci. Math. Hungar., 12 (1977), pp. 81–88.

[6] Á. Elbert and A. Laforgia, *On the square of the zeros of Bessel functions*, SIAM J. Math. Anal., 15 (1984), pp. 206–212.

[7] ———, *Further results on the zeros of Bessel functions*, Analysis, 5 (1985), pp. 71–86.

[8] M. E. H. Ismail and M. E. Muldoon, *On the variation with respect to a parameter of zeros of Bessel and q-Bessel functions*, J. Math. Anal. Appl., 135 (1988), pp. 187–207.

[9] E. Jahnke, F. Emde, and F. Lösch, *Tables of Higher Functions*, 6th ed., McGraw-Hill, New York, Toronto, London, 1960.

[10] A. Laforgia and M. E. Muldoon, *Inequalities and approximations for zeros of Bessel functions of small order*, SIAM J. Math. Anal., 14 (1983), pp. 383–388.

[11] L. Lorch and P. Szego, *Monotonicity of the difference of zeros of Bessel functions as a function of order*, Proc. Amer. Math. Soc., 15 (1964), pp. 91–96.

SIAM J. MATH. ANAL.
Vol. 24, No. 3, pp. 824–831, May 1993

018

SPECTRA OF JACOBI MATRICES, DIFFERENTIAL EQUATIONS ON THE CIRCLE, AND THE $su(1,1)$ LIE ALGEBRA*

JULIAN EDWARD†

Abstract. A family of differential operators on the circle is shown to be isospectral to a certain family of bilaterally infinite Jacobi matrices. The spectral properties of the differential operators are then used to explain a previously noted isospectral deformation of the Jacobi matrices. Differential operators on the circle are used to provide realizations of principle and complementary series representations of $su(1,1)$.

Key words. Jacobi matrices, spectrum, differential equations, Lie algebra, $su(1,1)$

AMS subject classifications. 17B15, 39A70

1. Introduction. A bilaterally infinite Jacobi matrix is a symmetric tridiagonal matrix of the form

$$M = \begin{pmatrix} \ddots & \ddots & b_0 & & & 0 \\ & b_0 & a_0 & b_1 & & \\ & & b_1 & a_1 & b_2 & \\ & & & b_2 & a_2 & b_3 \\ 0 & & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{1.1}$$

with $b_j > 0$ and $a_j \in \mathbb{R}$.

In cases where M, acting on $\ell^2(\mathbb{Z})$ (the set of square summable bilateral sequences), has discrete spectrum the eigenvalues have rarely been explicitly computed. One case where the eigenvalues have been computed explicitly is the case where M is connected with associated Meixner polynomials.

More explicitly, suppose

$$a_j = \delta j \quad \text{and} \quad b_j^2 = \alpha j^2 + \beta j + \gamma, \tag{1.2}$$

with $\alpha, \beta, \gamma, \delta \in \mathbb{R}$, and $\delta^2 > 4\alpha > 0$.

Then the spectrum of M was shown by Masson and Repka in [1] to be the following discrete set:

$$\lambda_j = j\sqrt{\delta^2 - 4\alpha} + \frac{1}{2}\left(1 + \frac{\beta}{\alpha}\right)(\sqrt{\delta^2 - 4\alpha} - \delta), \qquad j \in \mathbb{Z}. \tag{1.3}$$

To obtain this result, the authors above studied the properties of the left and right subdominant solutions to the second-order difference equation associated with M. Unfortunately, using these methods the authors were unable to explain the curious fact that the spectrum of M was independent of γ.

The main purpose of this note is to provide an alternative proof of (1.3) in a setting in which the independence of λ_j on γ is easily explained. It is also possible that the techniques used here could be useful in studying the spectra of other Jacobi matrices.

* Received by the editors March 17, 1992; accepted for publication (in revised form) September 3, 1992. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

† Department of Mathematics, University of Toronto, Toronto, Canada, M5S 1A1.

The key observation is the following. Let W be a closed curve, and let an arclength parametrization of W define a local coordinate x. Then the matrix M, acting on $\ell^2(\mathbb{Z})$, is isospectral to the following differential operator on W:

$$-\frac{i\partial}{\partial x}+q.$$

Here $q \in C^\infty(W)$, and the differential operator acts on the square integrable functions on W, $L^2(W)$. The observation is proven in §3.

The isospectral variation noted above will then be explained by the following.

LEMMA 1.1. *Suppose $\int q(x)dx = 0$. Then the spectrum of $-i\partial/\partial x + \gamma q(x)$ is independent of γ.*

This will be proven in §2. In §4, we use the connection between Jacobi matrices and differential operators on the unit circle, S^1, to obtain the principle and complementary series representations of the Lie algebra $su(1,1)$. The principle series representation is shown to be realized by a set of differential operators acting on $L^2(S^1)$. The complementary series representation is shown to be realized by a set of differential operators living on S^1 and acting on a Hilbert subspace of some Sobolev space over S^1.

The principle and complementary series representations of $su(1,1)$ arise in a number of differential equations [2]–[6], and the simplicity of the representations given here might make them a good approach to studying the properties of the differential equations.

2. The spectrum of a differential operator on S^1.

Let W be a closed curve in $\mathbb{R}^2$ of arclength L, and let an arclength parametrization of W provide a local coordinate x on W.

We define $L^2(W)$, the set of complex valued, square integrable functions on W, as

$$L^2(W) = \left\{ f \,\middle|\, \int_{x=0}^{L} |f(x)|^2 dx < \infty \right\}.$$

$L^2(W)$ is a Hilbert space equipped with inner product

$$\langle f, g\rangle = \frac{1}{L}\int_{x=0}^{L} f(x)\bar{g}(x)dx$$

and norm

$$\|f\|^2 = \frac{1}{L}\int_{x=0}^{L} |f(x)|^2 dx.$$

Let $C^\infty(W)$ be the set of infinitely differentiable functions on W.

Let $i = \sqrt{-1}$, and denote the differential operator $-i\partial/\partial x$ by D_x. Denote the spectrum of an operator A by $\sigma(A)$.

LEMMA 2.1. *Let $q \in C^\infty(W)$. Then*

$$\sigma(D_x + q) = \left\{ \frac{2\pi}{L}k + \frac{1}{L}\int_0^L q(x)\, dx;\ k \in \mathbb{Z} \right\}.$$

Proof. Suppose $(D_x+q)\phi = \lambda\phi$, $\phi \in L^2(W)$. Let $\phi(0) = C$. Then by uniqueness,

$$\phi(x_0) \;=\; C\,\exp\left(i\int_0^{x_0}(\lambda - q(x))dx\right).$$

Since $D_x + q - \lambda$ is an elliptic operator, it follows that $\phi \in C^\infty(W)$ [8]. In particular,

$$\phi(0) \;= \phi(L) \;=\; C\exp\left(i\int_0^{L}(\lambda - q(x))dx\right).$$

It follows that

$$\lambda \;=\; \frac{2\pi}{L}k + \frac{1}{L}\int_0^L q(x)dx \quad \text{with } k \in \mathbb{Z}.$$

Conversely, suppose $\lambda = 2\pi k/L + 1/L\int_0^L q(x)dx$. Let $\rho : \mathbb{R} \to W$ be an arclength preserving covering map. Define $\tilde{\phi} \in C^\infty(\mathbb{R})$ by

$$\tilde{\phi}(y_0) \;=\; \exp\left(i\int_0^{y_0}(\lambda - q\circ\rho(y))dy\right).$$

Then it is easy to verify that $\tilde{\phi}$ satisfies

$$\tilde{\phi}(y_0) = \tilde{\phi}(y_0 + L),$$
$$\left(-i\frac{\partial}{\partial y} + q\circ\rho\right)\tilde{\phi} = \lambda\tilde{\phi}.$$

It follows that there exists $\phi \in C^\infty(W)$, with $\phi\circ\rho = \tilde{\phi}$, such that

$$(D_x + q)\phi \;=\; \lambda\phi. \qquad \square$$

Note that Lemma 1 now follows.

Now let the usual arclength parametrization of S^1 provide the local coordinate θ.

THEOREM 1. *Let $A = (d + 2\sqrt{a}\cos\theta)D_\theta + 2e\sin\theta + 2e'\cos\theta$, with $d^2 > 4a$ and $e, e' \in \mathbb{R}$. Then $\sigma(A) = \{\sqrt{d^2 - 4a}\; k + e'(\sqrt{d^2-4a} - d)/\sqrt{a};\; k \in \mathbb{Z}\}$.*

Proof. Let $f : S^1 \to W$ be a diffeomorphism such that with respect to the coordinates θ and x on S^1 and W, respectively,

$$f'(\theta) \;=\; \frac{1}{d + 2\sqrt{a}\cos\theta}.$$

We use f to push A forward onto W. Let $u \in C^\infty(W)$. Then

$$\begin{aligned}((f_*A)u)(x) &= (A(u\circ f))(f^{-1}(x))\\ &= (D_x + q)u(x),\end{aligned}$$

where $q(f(x)) = 2e\sin\theta + 2e'\cos\theta$.

Since f_*A is isospectral to A, we can apply Lemma 2.

It follows from [7, eq. 3.645] that

$$\int_0^{2\pi}\frac{1}{d + 2\sqrt{a}\cos\theta}d\theta \;=\; \frac{2\pi}{\sqrt{d^2 - 4a}}. \tag{2.1}$$

Noting that $L = \int_0^{2\pi} f'(\theta)d\theta$, it follows from (2.1) that

$$\frac{1}{L}\int_0^L q(x)dx = \frac{1}{L}\int_0^{2\pi} \frac{(2e\sin\theta + 2e'\cos\theta)}{(d+2\sqrt{a}\cos\theta)d\theta}$$
$$= \left(\frac{e'}{\sqrt{a}}\right)(\sqrt{d^2-4a}-d). \qquad \square$$

The theorem now follows.

3. Differential operators on S^1 and Jacobi matrices. Let e_j be the column vector defined by

$$e_j = (\ldots,0,0,\underbrace{1}_{j},0,0,\ldots)^t.$$

The set of square summable bilateral sequences, denoted $\ell^2(\mathbb{Z})$, is given by

$$\left\{ u = \sum_{j=-\infty}^{\infty} u_j e_j \,\middle|\, u_j \in \mathbb{C}, \quad \sum_{j=-\infty}^{\infty} |u_j|^2 < \infty \right\}.$$

The inner product on $\ell^2(\mathbb{Z})$ of vectors $u = \sum_{j=-\infty}^{\infty} u_j e_j$ and $v = \sum_{j=-\infty}^{\infty} v_j e_j$ is given by

$$(u,v) = \sum_{j=-\infty}^{\infty} u_j \bar{v}_j.$$

PROPOSITION 1. *Let $M = (M_{m,n})$ be a bilaterally infinite Jacobi matrix acting on $\ell^2(\mathbb{Z})$ with $M_{n,n} = \delta n$, $M_{n,n-1} = M_{n-1,n} = \sqrt{\alpha n^2 + \beta n + \gamma}$, and $\delta^2 > 4\alpha > 0$. Then M has discrete spectrum.*

Proof. We prove this result by comparing M to a differential operator acting on $L^2(S^1)$. Let

$$B = (\delta + 2\sqrt{\alpha}\cos\theta)D_\theta + \sqrt{\alpha}e^{i\theta}.$$

Then B is a symmetric differential operator on $L^2(S^1)$, and, since $\delta + 2\sqrt{\alpha}\cos\theta > 0$, B is also elliptic. Denote its selfadjoint extension again by B. By the basic theory of pseudodifferential operators on compact manifolds, B has a discrete spectrum and $(B+i)^{-1}$ is a compact operator [8].

$L^2(S^1)$ is identified with $\ell^2(\mathbb{Z})$ by the unitary transformation $e^{ij\theta} \to e_j$. Via this identification, B induces the operator B' on $\ell^2(\mathbb{Z})$, and

$$(M-B')e_j = \left(\sqrt{\alpha(j+1)^2 + \beta(j+1) + \gamma} - \sqrt{\alpha}(j+1)\right)$$
$$\cdot e_{j+1} + \left(\sqrt{\alpha j^2 + \beta j + \gamma} - \sqrt{\alpha}j\right) e_{j-1}.$$

It follows that $M - B'$ is a bounded operator on $\ell^2(\mathbb{Z})$.

Since $(B'+i)^{-1}$ is a compact operator on $\ell^2(\mathbb{Z})$, it follows that $(M-B')(B'+i)^{-1}$ is also compact. Thus M differs from B' by a relatively compact perturbation, and it follows that M and B' have the same essential spectrum [9]. $\square$

Define a function on $\mathbb{Z}$ by

$$\pi_n = \begin{cases} \prod_{j=1}^{n}(\sqrt{a}j - e + e')/(\sqrt{a}(j-1) + e + e') & \text{if } n > 0, \\ 1 & \text{if } n = 0, \\ \prod_{j=1}^{-n}(-\sqrt{a}j + e + e')/(\sqrt{a}(1-j) - e + e') & \text{if } n < 0, \end{cases}$$

with $a, e, e' \in \mathbb{C}$.

LEMMA 3.1. *There exist constants $K_1, K_2 > 0$ such that*

$$K_1|n|^{(1-2e/\sqrt{a})3/4} \le |\pi_n| \le K_2|n|^{(1-2e/\sqrt{a})5/4} \quad \textit{if } n \neq 0.$$

Proof. Suppose that $n > 0$. Choose $j_0 \in \mathbb{N}$ such that for $j \ge j_0$, the following conditions are satisfied:

(i) $\sqrt{a}(j-1) + e + e' > 0$;

(ii) $|(\sqrt{a} - 2e)/(\sqrt{a}(j-1) + e + e')| < 1$;

(iii) $\frac{3}{4} < \sum_{\ell=0}^{\infty}(-1)^{\ell}/(\ell+1)((\sqrt{a} - 2e)/(\sqrt{a}(j-1) + e + e'))^{\ell} < \frac{5}{4}$.

Then, using the Taylor series of $\ln(1+x)$ for $|x| < 1$, we obtain

$$C_1 + \frac{3}{4}(\sqrt{a} - 2e)\sum_{j=j_0}^{n}\frac{1}{\sqrt{a}(j-1) + e + e'} \le \ln|\pi_n|$$

$$\le C_2 + \frac{5}{4}(\sqrt{a} - 2e)\sum_{j=j_0}^{n}\frac{1}{\sqrt{a}(j-1) + e + e'}$$

for some $C_1, C_2 > 0$. Hence

$$C_1' + \tfrac{3}{4}(1 - 2e/\sqrt{a})\ln n \le \ln|\pi_n| \le C_2' + \tfrac{5}{4}(1 - 2e/\sqrt{a})\ln n$$

for $C_1', C_2' > 0$.

Thus we obtain

$$e^{C_1'}n^{(1-2e/\sqrt{a})3/4} \le |\pi_n| \le e^{C_2'}n^{(1-2e/\sqrt{a})5/4}.$$

The proof for $n < 0$ is similar. □

We now prove (1.3).

THEOREM 2. *Let M satisfy the assumptions of Proposition* 1. *Then*

$$\sigma(M) = \left\{j\sqrt{\delta^2 - 4\alpha} + \frac{1}{2}\left(1 + \frac{\beta}{\alpha}\right)(\sqrt{\delta^2 - 4\alpha} - \delta);\ j \in \mathbb{Z}\right\}.$$

Proof. Consider the differential operator A on S^1 given by

$$A = (\delta + 2\sqrt{\alpha}\cos\theta)D_\theta + 2ei\sin\theta + 2e'\cos\theta,$$

with $e, e' \in \mathbb{C}$ chosen so that

$$2e'\sqrt{\alpha} - \alpha = \beta,$$
$$(e')^2 - e^2 + \sqrt{\alpha}e - \sqrt{\alpha}e' = \gamma. \quad □$$

We will show that A acting on $L^2(S^1)$ has the same spectrum as M acting on $\ell^2(\mathbb{Z})$. The theorem will then follow from Theorem 1. Note that by Theorem 1 and

Proposition 1 the operators A and M both have discrete spectrum; so it suffices to show that the eigenvalues coincide.

Under the identification $L^2(S^1) \sim \ell^2(\mathbb{Z}), A$ induces an operator A' on $\ell^2(\mathbb{Z})$ with

$$A'e_j = (\sqrt{\alpha}j + e + e')e_{j+1} + \delta j e_j + (\sqrt{\alpha}j - e + e')e_{j-1}.$$

Consider the operator $T : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$ given by $Te_j = 1/\sqrt{\pi_j}e_j$. Then

$$T^{-1}A'Te_j = \sqrt{\alpha(j+1)^2 + (2e'\sqrt{\alpha} - \alpha)(j+1) + (e')^2 - e^2 + \sqrt{\alpha}e - \sqrt{\alpha}e'}e_{j+1}$$
$$+\delta j e_j + \sqrt{\alpha j^2 + (2e'\sqrt{\alpha} - \alpha)j + (e')^2 - e^2 + \sqrt{\alpha}e - \sqrt{\alpha}e'}e_{j-1}.$$

Thus $M = T^{-1}A'T$. It remains to show that the eigenvalues of A and $T^{-1}A'T$ coincide.

Suppose $\lambda \in \sigma(A)$. Thus there exists $u(\theta) = \sum_{j=-\infty}^{\infty} u_j e^{ij\theta}$ such that

$$u \in L^2(S^1),$$
$$(A - \lambda)u = 0.$$

Since $A - \lambda$ is an elliptic differential operator, it follows that $u \in C^\infty(S^1)$. Hence

$$\lim_{|j|\to\infty} |u_j|\ |j|^p = 0 \quad \forall p \in \mathbb{N}. \tag{3.1}$$

Let $u' = \sum_{j=-\infty}^{\infty} u_j e_j$. It follows from Lemma 3 and (3.1) that

$$T^{-1}u' \in \ell^2(\mathbb{Z}).$$

Thus $T^{-1}u'$ is an eigenvector of $T^{-1}A'T$ with eigenvalue λ.

Conversely, suppose $\lambda \in \sigma(T^{-1}A'T)$. Thus there exists $u' = \sum u_j e_j$ satisfying $T^{-1}A'Tu' = \lambda u'$ with $u' \in \ell^2(\mathbb{Z})$. Consider the distribution on S^1 defined by

$$u(\theta) = \sum_{j=-\infty}^{\infty} \frac{u_j}{\sqrt{\pi_j}} e^{ij\theta}.$$

It is easily verified that $(A - \lambda)u = 0$ in the weak sense. Since $A - \lambda$ is elliptic, it follows that $(A - \lambda)u = 0$ in the strong sense and $u \in C^\infty(S^1)$. Thus $\lambda \in \sigma(A)$.

4. Representations of *su*(1,1). Consider the following differential operators on S^1:

$$T_1 = \cos\theta D_\theta + ae^{i\theta} + be^{-i\theta},$$
$$T_2 = \sin\theta D_\theta - aie^{i\theta} + bie^{-i\theta},$$
$$T_3 = D_\theta + a + b,$$

with $a, b \in \mathbb{C}$. The following is easily verified.

LEMMA 4.1. *$[T_1, T_2] = -iT_3$, $[T_2, T_3] = iT_1$, and $[T_3, T_1] = iT_2$.*

Thus for each $a, b \in \mathbb{C}$, the operators T_1, T_2, and T_3 form a representation on $su(1,1)$ on $C^\infty(S^1)$. We will show that for various values of a and b these operators are (up to conjugacy) the principle series and the complementary series.

Principle Series. Let $a = \frac{1}{2} - \sigma i$ and $b = \sigma i$, with $\sigma \in \mathbb{R}$, and consider T_1, T_2, T_3 acting on $L^2(S^1)$. As in the previous section, identify $L^2(S^1)$ with $\ell^2(\mathbb{Z})$ via the mapping $e^{ij\theta} \to e_j$, and let T_1', T_2' and T_3' be the transformed operators.

Thus

$$\begin{aligned}
T_1' ej &= (\tfrac{1}{2}j + \tfrac{1}{2} - \sigma i)e_{j+1} + (\tfrac{1}{2}j + \sigma i)e_{j-1},\\
T_2' ej &= -i(\tfrac{1}{2}j + \tfrac{1}{2} - \sigma i)e_{j+1} + i(\tfrac{1}{2}j + \sigma i)e_{j-1},\\
T_3' ej &= (j + \tfrac{1}{2})e_j.
\end{aligned}$$

Let π_n be the following complex valued function on $\mathbb{Z}$:

$$\pi_n = \begin{cases} \prod_{j=1}^{n} \left(\frac{j}{2} - \sigma i\right) \Big/ \left(\frac{j}{2} + \sigma i\right) & \text{if } n > 0,\\ 1 & \text{if } n = 0,\\ \prod_{j=1}^{-n} \left(\frac{1}{2} - \frac{j}{2} + \sigma i\right) \Big/ \left(\frac{1}{2} - \frac{j}{2} - \sigma i\right) & \text{if } n < 0. \end{cases}$$

Let $\mathcal{U} : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$ be the unitary transformation defined by $\mathcal{U}e_j = \sqrt{\pi_j}e_j$. Then it is easily verified that

$$\begin{aligned}
(\mathcal{U}T_1\mathcal{U}^{-1})e_j &= \sqrt{\tfrac{1}{4}(j+1)^2 + \sigma^2}\; e_{j+1} + \sqrt{\tfrac{1}{4}j^2 + \sigma^2}\; e_{j-1},\\
(\mathcal{U}T_2\mathcal{U}^{-1})e_j &= \tfrac{1}{i}\sqrt{\tfrac{1}{4}(j+1)^2 + \sigma^2}\; e_{j+1} - \tfrac{1}{i}\sqrt{\tfrac{1}{4}j^2 + \sigma^2}\; e_{j-1},\\
(\mathcal{U}T_3\mathcal{U}^{-1})e_j &= (j + \tfrac{1}{2})\; e_j.
\end{aligned}$$

This is precisely the action of the principal series representation.

Complementary series. Let $a = (t-s)/2$ and $b = (t+s)/2$, with $-1+|t| < s < -|t|$. Define a function π_n on $\mathbb{Z}$ by

$$\pi_n = \begin{cases} \prod_{j=1}^{n}(j+t+s)/(j-1+t-s) & \text{if } n > 0,\\ 1 & \text{if } n = 0,\\ \prod_{j=1}^{-n}(-j+t-s)/(1-j+t+s) & \text{if } n < 0. \end{cases}$$

By Lemma 3, either π_n or π_n^{-1} will be an unbounded function of n. Thus the technique used to obtain the principle series is not applicable in this case.

Instead, adapting the methods used in [10], we consider the weighted ℓ^2 space $\ell^2(\mathbb{Z}, \pi_j)$, with

$$\ell^2(\mathbb{Z}, \pi_j) = \left\{ \sum_{j=-\infty}^{\infty} u_j e_j \,\middle|\, u_j \in \mathbb{C} \quad \text{and} \quad \sum_{j=-\infty}^{\infty} |u_j|^2 \pi_j < \infty \right\}.$$

The inner product of $v = \sum_{j=-\infty}^{\infty} v_j e_j$ and $u = \sum_{j=-\infty}^{\infty} u_j e_j$ is given by $(u,v)_\pi = \sum_{j=-\infty}^{\infty} u_j \bar{v}_j \pi_j$.

Under the mapping $e_j \to e^{ij\theta}$, $\ell^2(\mathbb{Z}, \pi_j)$ is identified with some Hilbert space H, which, by Lemma 3, is the proper subspace of some Sobolev space.

Consider the operators T_1, T_2, and T_3 acting on H. These induce operators T_1', T_2', and T_3' acting on $\ell^2(\mathbb{Z}, \pi_j)$, given by

$$\begin{aligned}
T_1' e_j &= \left(\tfrac{1}{2}\right)(j + t - s)e_{j+1} + \left(\tfrac{1}{2}\right)(j + t + s)e_{j-1},\\
T_2' e_j &= \left(\tfrac{1}{2i}\right)(j + t - s)e_{j+1} - \left(\tfrac{1}{2i}\right)(j + t + s)e_{j-1},\\
T_3' e_j &= (j + t)e_j.
\end{aligned}$$

Let $\dot{e}_j = e_j/\sqrt{\pi_j}$. Then $\{\dot{e}_j | j \in \mathbb{Z}\}$ provides an orthonormal basis of $\ell^2(\mathbb{Z}, \pi_j)$. It is easily verified that

$$(4.1) \quad \begin{cases} (T_1'\dot{e}_{j+1}, \dot{e}_j)_\pi = (\dot{e}_{j+1}, T_1'\dot{e}_j)_\pi = \frac{1}{2}\sqrt{(j+t-s)(j+1+t+s)}, \\ (T_1'\dot{e}_{j+n}, \dot{e}_j)_\pi = 0 \quad \text{for } n \neq -1, 1. \end{cases}$$

Let $\mathcal{U} : \ell^2(\mathbb{Z}, \pi_n) \to \ell^2(\mathbb{Z})$ be the unitary transformation given by $\mathcal{U}\dot{e}_j = e_j$. Then it follows from (4.1) that

$$\mathcal{U}T_1'\mathcal{U}^{-1}e_j = \tfrac{1}{2}\sqrt{(j+t-s)(j+1+t+s)}e_{j+1} + \tfrac{1}{2}\sqrt{(j-1+t-s)(j+t+s)}e_{j-1}.$$

Similarly,

$$\mathcal{U}T_2'\mathcal{U}^{-1}e_j = \tfrac{1}{2i}\sqrt{(j+t-s)(j+1+t+s)}e_{j+1} - \tfrac{1}{2i}\sqrt{(j-1+t-s)(j+t+s)}e_{j-1},$$

and

$$\mathcal{U}T_3'\mathcal{U}^{-1}e_j = (j+t)e_j.$$

Thus the operators $\mathcal{U}T_1'\mathcal{U}^{-1}$, $\mathcal{U}T_2'\mathcal{U}^{-1}$, and $UT_3'\mathcal{U}^{-1}$ give the action of the complementary series representation of $su(1,1)$ on $\ell^2(\mathbb{Z})$.

REFERENCES

[1] D. Masson and J. Repka, *Spectral theory of Jacobi matrices in $l^2(\mathbb{Z})$ and the $su(1,1)$ Lie algebra*, SIAM J. Math. Anal., 22 (1991), pp. 1131–1146.

[2] Y. Alhassid, F. Gursey, and F. Iachello, *Group theory approach to scattering*, Ann. Phys., 148 (1983), pp. 346–380.

[3] A. Barut, A. Inomata, and R. Wilson, *The generalised morse oscillator in the $SO(4,2)$ dynamical group scheme*, J. Math. Phys., 28 (1987), pp. 605–611.

[4] D. Masson, *Schrodinger's Equation and Continued Fractions*, Internat. J. Quantum Chem., Quantum Chemistry Symposium, 21 (1987), pp. 699–712.

[5] W. Miller, *On Lie algebras and some special functions of mathematical physics*, Mem. Amer. Math. Soc., 50 (1964), pp. i–43.

[6] P. Ojha, *$SO(2,1)$ Lie algebra and the Jacobi-Matrix method for scattering*, Phys. Rev. A, 34 (1986), pp. 969–977.

[7] I. Gradshteyn and I. Ryzhik, *Tables of Integrals, Series and Products*, Academic Press, New York, 1967.

[8] M. Shubin, *Pseudodifferential Operators and Spectral Theory*, Springer-Verlag, Berlin, 1985.

[9] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, Vol.* IV, Academic Press, San Diego, CA, 1978.

[10] M. Ismail, J. Letessier, D. Masson, and G. Valent, *Birth and death processes and orthogonal polynomials*, in Proc. NATO Advanced Institute on Orthogonal Polynomials and their Applications, Columbus, Ohio, 1989; Orthogonal Polynomials: Theory and Practice, P. Nevai, ed., Kluwer, Dordrecht, the Netherlands, 1990, pp. 229–255.

SIAM J. MATH. ANAL.
Vol. 24, No. 4, 833–839, July 1993

CONVEXITY OF SOLUTIONS TO SOME ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS*

ANTONIO GRECO† AND GIOVANNI PORRU†

Abstract. This paper is concerned with positive solutions $u(x)$ to some special elliptic partial differential equations in a bounded convex domain $\Omega \subset R^N$. For a decreasing function $g(t)$ the transformation $v = g(u)$ is performed and, under appropriate restrictions on $g(t)$, it is proved that $v(x)$ is convex in Ω. Consequently, the level sets of $u(x)$ are convex.

Key words. elliptic equations, maximum principles, convexity

AMS subject classifications. 35B50, 35J99

1. Introduction.

In [1] Brascamp and Lieb show that the first (positive) eigenfunction of the Laplacian in a strictly convex domain $\Omega \subset R^N$ is log-concave. Their method makes use of the fact that the linear parabolic operator $\partial/\partial t - \Delta$, under homogeneous Dirichlet boundary conditions, preserves log-concavity of the initial data. Extending this method, Lions in [12] proves the log-concavity for the solution $u > 0$ of the problem

$$\Delta u = -h(u)u \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega,$$

where $h = h(t)$ is a positive real function satisfying

$$h'(t) \leq 0, \qquad h''(t)t + h'(t) \leq 0.$$

In [9], Korevaar gives a new proof of the above result by considering the concavity function

$$C(v; x, y) = v(z) - \frac{[v(x) + v(y)]}{2}, \qquad z = \frac{(x+y)}{2}, \tag{1.1}$$

where $x, y \in \Omega$. Of course, $v(x)$ is convex in Ω if and only if $C(v; x, y) \leq 0$ for all $(x, y) \in \Omega \times \Omega$. Consider the following elliptic equation:

$$a^{ij}(Dv)v_{ij} = b(x, v, Dv) \quad \text{in } \Omega, \tag{1.2}$$

where $v_i = \partial v/\partial x_i$, Dv is the gradient of v and the summation convention (here and in the sequel) over repeated indices is in effect. In [9] Korevaar shows that if $v(x)$ satisfies (1.2), where $b(x, v, p)$ is increasing with respect to v and jointly concave with respect to (x, v), then the concavity function (1.1) cannot attain any positive maximum in $\Omega \times \Omega$.

Extensions of Korevaar's method have been found independently by Caffarelli and Friedman [2], Caffarelli and Spruck [3], Kawohl [5], Kennington [6]. In [5] and [6] the maximum principle for the function (1.1) is proved under the conditions: $b \geq 0$, b strictly increasing with respect to v and $b(x, v, p)$ harmonic concave with respect to (x, v). (Note that if b is positive and concave then it is harmonic concave; the reverse

*Received by the editors January 17, 1992; accepted for publication (in revised form) October 19, 1992.

†Dipartimento di Matematica, Via Ospedale 72, 09124 Cagliari, Italy.

is not true in general.) The method of proof is by contradiction. By assuming that the function (1.1) has a positive maximum in $\Omega \times \Omega$ and by a suitable use of (1.2) it is possible to reach a contradiction.

In this paper we present an alternative proof of the last result by finding an elliptic (degenerate) inequality (in R^{2N}) satisfied by $C(v; x, y)$. Our proof does not require the assumption $b \geq 0$ and applies when $\partial b/\partial v \geq 0$. We also propose a new proof of a boundary point lemma.

Other techniques have been developed for studying convexity properties of solutions in [10] for ring domains and in [7] for parabolic equations.

2. A maximum principle. Let us state an unusual form of the classical maximum principle for elliptic degenerate inequalities.

LEMMA 2.1. *Let G be a domain in R^m and ϕ be a $C^2(G)$ solution of the differential inequality*

$$b^{hk}\phi_{hk} + b^h\phi_h \geq c\phi, \tag{2.1}$$

where $c \geq 0$ and the $m \times m$ matrix $[b^{hk}]$ is positive semidefinite in G. If for any compact set $K \subset G$ there exists a real number M and ν $(1 \leq \nu \leq m)$ indices $h_1, \dots, h_\nu$ such that it results (at each point in K) in

$$Mb^{h_i h_i} + b^{h_i} \geq 0, \qquad i = 1, \dots, \nu, \tag{2.2}$$

and

$$\sum_{i=1}^{\nu}\left(Mb^{h_i h_i} + b^{h_i}\right) > 0, \tag{2.3}$$

then ϕ cannot have any local positive maximum in G.

Proof. Suppose there are a point x_0 and a compact set K with $x_0 \in K \subset G$ so that $\Phi(x_0) > 0$ and $\Phi(x_0) > \Phi(x)$ for all $x \in \partial K$. We can choose $\varepsilon > 0$ small enough so that the function

$$\psi(x) = \phi(x) + \varepsilon\left(\exp(Mx^{h_1}) + \cdots + \exp(Mx^{h_\nu})\right)$$

has a local positive maximum value at some point $x_1 \in K$. At x_1 it must be

$$b^{hk}\psi_{hk} + b^h\psi_h \leq 0. \tag{2.4}$$

On the other hand, by using (2.1), (2.2), and (2.3) we find (at each point in K):

$$b^{hk}\psi_{hk} + b^h\psi_h \geq c\Phi + \varepsilon M \sum_{i=1}^{\nu}(Mb^{h_i h_i} + b^{h_i})\exp(Mx^{h_i}) > c\Phi.$$

The last inequality contradicts (2.4) at the point x_1. The lemma is proved. For $\nu = 1$ this lemma is proved in [4, pp. 31–32].

We also need the following result about matrices.

LEMMA 2.2. *Let $A = [a^{ij}]$ be a $N \times N$ positive semidefinite matrix. Then the $2N \times 2N$ new matrix*

$$B = \begin{pmatrix} \sigma^2 A & \sigma\tau A \\ \sigma\tau A & \tau^2 A \end{pmatrix}$$

is positive semidefinite for any pair of real numbers σ, τ.

Proof. If ξ is a vector in R^{2N}, let us denote by x the first N coordinates of ξ and by y the last N coordinates. We have $(B\xi, \xi) = (Az, z)$, where $z = \sigma x + \tau y$. The lemma follows.

DEFINITION 2.1. A function $f(x)$ is said to be harmonic concave in a convex domain $G \subset R^m$ if for any pair $x_1, x_2 \in G$ for which $f(x_1) + f(x_2) > 0$ it results

$$[f(x_1) + f(x_2)]f((x_1 + x_2)/2) \geq 2f(x_1)f(x_2).$$

Since no condition is imposed when $f(x_1) + f(x_2) \leq 0$ any nonpositive function is (according to our definition) harmonic concave. Of course, if a function is concave, then it is harmonic concave. It is easy to prove that if $f(x)$ is harmonic concave, then so is $f(x) + c$ for any negative constant c. (For a positive constant c the previous result is not true in general.) Definition 2.1 is natural for proving the following.

THEOREM 2.1. *Let Ω be a convex domain in R^N and let $v \in C^2(\Omega)$ be a solution of the strictly elliptic equation*

$$a^{ij}(Dv)v_{ij} = b(x, v, Dv), \tag{2.5}$$

where $a^{ij}(Dv)$ are smooth and $b(x, v, Dv)$ is smooth, nondecreasing with respect to v and harmonic concave with respect to (x, v). Then the concavity function defined in (1.1) *cannot have any local positive maximum in $\Omega \times \Omega$.*

Proof. We may suppose $a^{ij} = a^{ji}$. If v is a solution of (2.5) we set

$$\phi(x, y) = 2v(z) - v(x) - v(y), \tag{2.6}$$

where (here and in the sequel),

$$z = \frac{(x + y)}{2}.$$

Of course, we have $\phi(x, y) = 2C(v; x, y)$, with $C(v; x, y)$ introduced in (1.1). Let $(x, y) \in \Omega \times \Omega$ be a point at which

$$b(x, v(x), Dv(z)) + b(y, v(y), Dv(z)) > 0. \tag{2.7}$$

Since b is harmonic concave with respect to (x, v) we have

$$\begin{aligned} &[b(x, v(x), p) + b(y, v(y), p)]b\left(z, \frac{v(x) + v(y)}{2}, p\right) \\ &\qquad -2b(x, v(x), p)b(y, v(y), p) \geq 0 \end{aligned} \tag{2.8}$$

for all $p \in R^N$. At each point in $\Omega \times \Omega$ where (2.7) holds we define the second-order operator

$$\begin{aligned} L\phi \equiv &b^2(y, v(y), Dv(z))a^{ij}(Dv(z))\phi_{x_i x_j} \\ &+ 2b(y, v(y), Dv(z))b(x, v(x), Dv(z))a^{ij}(Dv(z))\phi_{x_i y_j} \\ &+ b^2(x, v(x), Dv(z))a^{ij}(Dv(z))\phi_{y_i y_j}. \end{aligned}$$

By Lemma 2.2, the operator L is elliptic (degenerate). By using (2.6) we find

$$\begin{aligned} L\phi = &b^2(y, v(y), Dv(z))[2^{-1}a^{ij}(Dv(z))v_{z_i z_j} - a^{ij}(Dv(z))v_{x_i x_j}] \\ &+ 2b(y, v(y), Dv(z))b(x, v(x), Dv(z))2^{-1}a^{ij}(Dv(z))v_{z_i z_j} \\ &+ b^2(x, v(x), Dv(z))[2^{-1}a^{ij}(Dv(z))v_{z_i z_j} - a^{ij}(Dv(z))v_{y_i y_j}]. \end{aligned}$$

Since

$$\phi_{x_s}(x,y) = v_{z_s}(z) - v_{x_s}(x), \qquad \phi_{y_s}(x,y) = v_{z_s}(z) - v_{y_s}(y),$$

we have

$$a^{ij}(Dv(z))v_{x_ix_j} = a^{ij}(Dv(x))v_{x_ix_j} + a^{ij}_{p_s} v_{x_ix_j}\phi_{x_s}.$$

Here and below we frequently apply the mean value theorem, so that in this instance $a^{ij}_{p_s}$ is evaluated at an intermediate point between $Dv(z)$ and $Dv(x)$. Similarly we find

$$a^{ij}(Dv(z))v_{y_iy_j} = a^{ij}(Dv(y))v_{y_iy_j} + a^{ij}_{p_s} v_{y_iy_j}\phi_{y_s}.$$

After use of these substitutions and of (2.5) at the points z, x, and y we find

$$\begin{aligned} L\phi{+}b^2&(y,v(y),Dv(z))a^{ij}_{p_s}v_{x_ix_j}\phi_{x_s} + b^2(x,v(x),Dv(z))a^{ij}_{p_s}v_{y_iy_j}\phi_{y_s} \\ &= 2^{-1}[b(y,v(y),Dv(z)) + b(x,v(x),Dv(z))]^2 b(z,v(z),Dv(z)) \\ &\quad - b^2(y,v(y),Dv(z))b(x,v(x),Dv(x)) - b^2(x,v(x),Dv(z))b(y,v(y),Dv(y)). \end{aligned}$$

We also have

$$\begin{aligned} b(z,v(z),Dv(z)) &= b\left(z, \frac{v(x)+v(y)}{2}, Dv(z)\right) + \frac{b_v\phi}{2}, \\ b(x,v(x),Dv(x)) &= b(x,v(x),Dv(z)) - b_{p_s}\phi_{x_s}, \\ b(y,v(y),Dv(y)) &= b(y,v(y),Dv(z)) - b_{p_s}\phi_{y_s}. \end{aligned}$$

Hence we find

$$\begin{aligned} L\phi{+}b^2&(y,v(y),Dv(z))Q^s\phi_{x_s} + b^2(x,v(x),Dv(z))P^s\phi_{y_s} \\ &= 4^{-1}[b(y,v(y),Dv(z)) + b(x,v(x),Dv(z))]^2 b_v\phi \\ &\quad + 2^{-1}\Bigg\{[b(x,v(x),p) + b(y,v(y),p)]^2 b\left(z, \frac{v(x)+v(y)}{2}, p\right) \\ &\qquad - 2[b(x,v(x),p) + b(y,v(y),p)]b(x,v(x),p)b(y,v(y),p)\Bigg\}, \end{aligned}$$

where $p = Dv(z)$ and

$$(2.9) \qquad Q^s = a^{ij}_{p_s}v_{x_ix_j} - b_{p_s}, \qquad P^s = a^{ij}_{p_s}v_{y_iy_j} - b_{p_s}.$$

Finally, by using hypothesis (2.8), we find

$$(2.10) \qquad L\phi + b^2(y,v(y),Dv(z))Q^s\phi_{x_s} + b^2(x,v(x),Dv(z))P^s\phi_{y_s} \geq c\phi,$$

where

$$c = 4^{-1}[b(y,v(y),Dv(z)) + b(x,v(x),Dv(z))]^2 b_v.$$

Since the entries in the first diagonal of the matrix relative to our operator L are

$$b^2(y,v(y),Dv(z))a^{ii}(Dv(z)) \quad \text{and} \quad b^2(x,v(x),Dv(z))a^{ii}(Dv(z)),$$

the assumptions of Lemma 2.1 are fulfilled. In fact, we may choose $m = 2N$, $h_1 = 1$, $h_2 = N+1$, and

$$M = 1 + \max\left[\sup\frac{|Q^1|}{a^{11}}, \sup\frac{|P^1|}{a^{11}}\right],$$

where the sup is taken over a compact set in $\Omega \times R^{N+1}$. Now we consider the points $(x, y) \in \Omega \times \Omega$ at which

$$b(x, v(x), Dv(z)) + b(y, v(y), Dv(z)) \leq 0. \tag{2.11}$$

At these points we define

$$L\phi \equiv a^{ij}(Dv(z))\phi_{x_i x_j} - 2a^{ij}(Dv(z))\phi_{x_i y_j} + a^{ij}(Dv(z))\phi_{y_i y_j}.$$

By Lemma 2.2, the operator L is elliptic (degenerate). By using (2.6) we find

$$\begin{aligned} L\phi =& 2^{-1}a^{ij}(Dv(z))v_{z_i z_j} - a^{ij}(Dv(z))v_{x_i x_j} \\ &- a^{ij}(Dv(z))v_{z_i z_j} + 2^{-1}a^{ij}(Dv(z))v_{z_i z_j} - a^{ij}(Dv(z))v_{y_i y_j}. \end{aligned}$$

By using the mean value theorem and equation (2.5) the last equality gives

$$\begin{aligned} L\phi &= - b(x, v(x), Dv(x)) - a^{ij}_{p_s} v_{x_i x_j} \phi_{x_s} - b(y, v(y), Dv(y)) - a^{ij}_{p_s} v_{y_i y_j} \phi_{y_s} \\ &= - b(x, v(x), Dv(z)) - Q^s \phi_{x_s} - b(y, v(y), Dv(z)) - P^s \phi_{y_s}, \end{aligned}$$

where Q^s and P^s are given by (2.9). The latter equation and assumption (2.11) imply

$$L\phi + Q^s \phi_{x_s} + P^s \phi_{y_s} \geq 0. \tag{2.12}$$

Hence the function $\phi(x, y)$ satisfies either (2.10) or (2.12) in $\Omega \times \Omega$. The theorem follows by Lemma 2.1.

Remark 2.1. Theorem 2.1 is proved (by using different methods) in [5] and in [6] in case of $b \geq 0$, $\partial b / \partial v > 0$, and in [9] in case b is concave.

3. Convexity results.

If $v(x)$ is not convex in Ω, then we have the following.

(i) $C(v; x, y)$ has a local positive maximum in $\Omega \times \Omega$; or

(ii) $C(v; x, y)$ becomes positive as (x, y) approaches $\partial(\Omega \times \Omega)$.

The (i) possibility has been discussed in Theorem 2.1. Now we examine the (ii) possibility.

LEMMA 3.1. *Let $\Omega \subset R^N$ be a bounded domain, strictly convex and with a smooth boundary $\partial\Omega$. Let $u \in C^1(\bar{\Omega})$, $u > 0$ in Ω, $u = 0$ on $\partial\Omega$ and $u_n < 0$ on $\partial\Omega$, where n is the exterior normal to $\partial\Omega$. Let $g : R^+ \longrightarrow R$ be a smooth function satisfying*

$$\lim_{u \to 0} g(u) = a \quad (\in R), \tag{3.1}$$

$$g'(u) < 0, \qquad \lim_{u \to 0} g'(u) = -\infty. \tag{3.2}$$

If $C(g(u); x, y)$ has not positive maxima in $\Omega \times \Omega$, then $C(g(u); x, y)$ cannot become positive as (x, y) approaches $\partial(\Omega \times \Omega)$.

Proof. We refer to [5, p. 115].

LEMMA 3.2. *Let $\Omega \subset R^N$ be a bounded domain, strictly convex and with a smooth boundary $\partial\Omega$. Let $u \in C^2(\bar{\Omega})$, $u > 0$ in Ω, $u = 0$ on $\partial\Omega$, and $u_n < 0$ on $\partial\Omega$, where n is the exterior normal to $\partial\Omega$. Let $g : R^+ \longrightarrow R$ be a smooth function satisfying*

$$\lim_{u \to 0} g(u) = +\infty, \tag{3.3}$$

$$g'(u) < 0, \qquad \lim_{u\to 0} g'(u) = -\infty, \tag{3.4}$$

$$g''(u) > 0, \qquad \lim_{u\to 0} g'(u)/g''(u) = 0. \tag{3.5}$$

Then $C(g(u); x, y)$ cannot become positive as (x, y) approaches $\partial(\Omega \times \Omega)$.

Proof. Otherwise there exist sequences $\{x_j\}$, $\{y_j\}$ in Ω with limits x_0, y_0 such that $(x_0, y_0) \in \partial(\Omega \times \Omega)$, and, for some positive ε,

$$C(g(u); x_j, y_j) \geq \varepsilon, \qquad j = 1, 2, \ldots. \tag{3.6}$$

By definition of C and (3.6) we have

$$g(u(z_j)) - 2^{-1}[g(u(x_j)) + g(u(y_j))] \geq \varepsilon, \qquad z_j = \frac{(x_j + y_j)}{2}. \tag{3.7}$$

If $x_0 \neq y_0$, then $(x_0 + y_0)/2 \in \Omega$. As j approaches infinity, $g(u(z_j))$ approaches $g(u((x_0 + y_0)/2))$ (a finite number), whereas x_j or y_j (or both) approaches $\partial\Omega$, where u vanishes. Hence, by using (3.3) we reach a contradiction. Suppose $x_0 = y_0$. Let us rewrite (3.7) as

$$[g(u(z_j)) - g(u(x_j))] + [g(u(z_j)) - g(u(y_j))] \geq 2\varepsilon.$$

By using the mean value theorem we find

$$g'(u(\xi_j))u_{\nu_j}(\xi_j)|x_j - y_j| - g'(u(\eta_j))u_{\nu_j}(\eta_j)|x_j - y_j| \geq 4\varepsilon,$$

where ξ_j lies between x_j and z_j, η_j lies between z_j and y_j, and ν_j is the unit vector $\nu_j = (y_j - x_j)/|y_j - x_j|$ (of course $x_j \neq y_j$ by (3.7)). By using once more the mean value theorem we obtain

$$-\Big(g''\big(u(\zeta_j)\big)u^2_{\nu_j}(\zeta_j) + g'\big(u(\zeta_j)\big)u_{\nu_j\nu_j}(\zeta_j)\Big)|x_j - y_j||\eta_j - \xi_j| \geq 4\varepsilon, \tag{3.8}$$

where ζ_j lies between ξ_j and η_j. Obviously ζ_j approaches x_0 as j approaches infinity. We may take subsequences $\{x_{j_k}\}$, $\{y_{j_k}\}$ so that the corresponding sequence of unit vectors $\{\nu_{j_k}\}$ converges to a unit vector ν. For short we suppose the original sequence to have this property. Let ν be tangential to $\partial\Omega$ at x_0. Since u vanishes and $u_n < 0$ on $\partial\Omega$ (by assumption), the Hessian matrix D^2u is strictly negative in all tangential directions. Consequently, since $g''(u) > 0$ and $g'(u) < 0$ as $j \longrightarrow \infty$ we get a contradiction in (3.8). Finally, let ν be nontangential. Then we have $u^2_\nu(x_0) > 0$. Assumptions (3.5) lead again to a contradiction (as $j \to \infty$) in (3.8). The lemma is proved.

Lemma 3.2 is stated in [5] without assumptions (3.5). The proof is similar to ours, but only the case $x_0 \neq y_0$ is taken into account. Lemma 3.1 and Lemma 3.2 are also proved in [9] by using assumptions (3.4), (3.5), and the additional restriction $\lim_{u\to 0} g(u)/g'(u) = 0$.

Acknowledgment. We are grateful to the referee for some suggestions that allowed us to improve our manuscript.

REFERENCES

[1] H. J. BRASCAMP AND E. H. LIEB, *On extensions of the Brunn–Minkowski and Prékopa–Leindler theorems, including inequalities for log-concave functions, and with an application to the diffusion equation*, J. Funct. Anal., 22 (1976), pp. 366–389.

[2] L. A. CAFFARELLI AND A. FRIEDMAN, *Convexity of solutions of semilinear elliptic equations*, Duke Math. J., 52 (1985), pp. 431–456.

[3] L. A. CAFFARELLI AND J. SPRUCK, *Convexity properties of solutions to some classical variational problems*, Comm. Partial Differential Equations, 7 (1982), pp. 1337–1379.

[4] D. GILBARG AND N.S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1983.

[5] B. KAWOHL, *Rearrangements and convexity of level sets in* PDE, Lecture Notes in Math. 1150, Springer-Verlag, New York, 1985.

[6] A. KENNINGTON, *An Improved Convexity Maximum Principle and Some Applications*, Ph.D. thesis, Univeristy of Adelaide, Adelaide, Australia, 1984.

[7] ———, *Convexity of level curves for an initial value problem*, J. Math. Anal. Appl., 133 (1988), pp. 324–330.

[8] N.J. KOREVAAR, *Capillarity surface convexity above convex domains*, Indiana Univ. Math. J., 32 (1983), pp. 73–82.

[9] ———, *Convex solutions to nonlinear elliptic and parabolic boundary value problems*, Indiana Univ. Math. J., 32 (1983), pp. 603–614.

[10] ———, *Convexity of level sets for solutions to elliptic ring problems,* Comm. Partial Differential Equations, 15 (1990), pp. 541–556.

[11] N. J. KOREVAAR AND J. L. LEWIS, *Convex solutions to certain elliptic equations have constant rank Hessian,* Arch. Rational Mech. Anal., 97 (1987), pp. 19–32.

[12] P. L. LIONS, *Two geometrical properties of solutions of semilinear problems*, Appl. Anal., 12 (1981), pp. 267–272.

[13] M. H. PROTTER AND H. F. WEINBERGER, *Maximum principles in differential equations*, Prentice Hall, Englewood Cliffs, NJ, 1967.

SIAM J. MATH. ANAL.
Vol. 24, No. 4, pp. 840–865, July 1993

002

ONE-PHASE RIEMANN PROBLEM AND WAVE INTERACTIONS IN SYSTEMS OF CONSERVATION LAWS OF MIXED TYPE*

HAITAO FAN[†]

Abstract. The Riemann problem for any Riemann data lying outside of the elliptic region of a system of conservation laws of mixed type is established. The approach is a vanishing viscosity one. The solutions constructed are also admissible by the traveling wave criterion. The structure of the solutions is analyzed. Interactions among phase boundaries and shocks are studied.

Key words. conservation laws, phase transitions, traveling wave criterion, vanishing viscosity method, Riemann problem, shocks

AMS subject classifications. primary 35L65; secondary 35L67, 76L05

1. Introduction.

The field equations which express the principles of balance for one-dimensional continuous media are typically the nonlinear systems of conservation laws

$$U_t + F(U)_x = 0. \tag{1.1}$$

If the Jacobian matrix

$$\frac{\partial F}{\partial U} \tag{1.2}$$

has n distinct real eigenvalues, the system (1.1) is called strictly hyperbolic. If (1.2) has real eigenvalues and some of them are equal to each other at some point $U_0 \in \mathbb{R}^n$, the system (1.1) is said to be nonstrictly hyperbolic type and the point U_0 is called an umbilic point of (1.1). Real problems are often not strictly hyperbolic. For these kinds of systems, transitional shocks must be introduced into the system and the geometry of the state space is singular (cf. [11]–[15], [24], [25], [29]). Antman [1] discussed some interesting phenomena in the governing system of equations of a special class of motion, part of a very general family of nonlinear viscoelastic materials of hyperbolic-parabolic type.

Some models, for example, Stone's model (cf. [11]), the model for an elastic bar or van der Waals fluids undergoing phase transitions etc., have elliptic regions in which the eigenvalues of (1.2) are complex. The presence of elliptic regions adds more complexity to the systems, for instance, the local analysis may not apply in these kinds of systems and the problems must be considered in the large.

To study the mathematical phenomena involved in nonstrictly and hyperbolic-elliptic mixed type systems of conservation laws, we consider the p-system

$$u_t + p(w)_x = 0, \tag{1.3a}$$

$$w_t - u_x = 0, \qquad x \in \mathbb{R}, t > 0, \tag{1.3b}$$

*Received by the editors March 26, 1992; accepted for publication (in revised form) September 10, 1992.

[†] Center for Dynamical Systems and Nonlinear Studies, School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

which describes the one-dimensional longitudinal isothermal motion in elastic bars or fluids, where u is the velocity, w the specific volume, and p the pressure. For our purposes we set the function p to satisfy $p \in C^1(\mathbb{R})$ and

$$\begin{aligned} p'(w) &< 0 \quad \text{for } w \notin [\alpha, \beta], \\ p'(w) &> 0 \quad \text{for } w \in (\alpha, \beta) \end{aligned} \tag{1.3c}$$

so as to make the system (1.3a), (1.3b) a hyperbolic-elliptic mixed type system. We call the region $w < \alpha$ the α-phase region and $w > \beta$ the β-phase region. A shock with two ends in different phase regions is called a phase boundary. See Fig. 1.

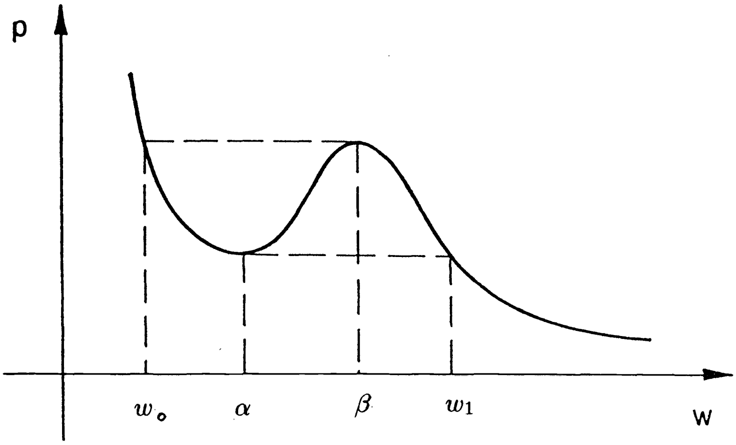


FIG. 1

In this paper, we establish the existence of the Riemann problem of (1.3) with arbitrary initial data

$$(u(x,0), w(x,0)) = \begin{cases} (u_-, w_-) & \text{for } x < 0, \\ (u_+, w_+) & \text{for } x > 0, \end{cases} \tag{1.4a}$$

$$w_\pm \notin (\alpha, \beta). \tag{1.4b}$$

We investigate the structure of the solutions we constructed. We also study the interactions among phase boundaries and "ordinary" shocks. We note that our technique can be generalized to the case where the system (1.1a), (1.1b) has finitely many elliptic regions as well as umbilic points.

Just as for hyperbolic systems of conservation laws, there is the issue of admissibility of solutions for the systems of hyperbolic-elliptic type like (1.3). Slemrod [24] proposed that an admissible solution of the Cauchy problem of (1.3) should be a $\epsilon \longrightarrow 0+$ limit of the solution (u_ϵ, w_ϵ) of the Cauchy problem

$$u_t + p(w)_x = \epsilon u_{xx} - \epsilon^2 A w_{xxx} - 2\epsilon^2 D w_x w_{xx}, \tag{1.5a}$$

$$w_t - u_x = 0, \tag{1.5b}$$

with the same initial value. We notice that Truskinovskii [30] independently proposed this criterion. A localized version of this criterion is the viscosity-capillarity traveling wave criterion, or the traveling wave criterion for short, which says that a shock of speed s connecting (u_1, w_1) and (u_2, w_2), satisfying Rankine–Hugoniot conditions, is admissible if (1.5) has a traveling wave solution of speed s connecting w_1 and w_2. When $A = \frac{1}{4}$ and $D = 0$, (1.5) reduces to the common form of viscosity

$$u_t + p(w)_x = \epsilon u_{xx}, \tag{1.6a}$$

$$w_t - u_x = \epsilon w_{xx}, \tag{1.6b}$$

with the same initial value. The corresponding traveling wave equation states that a shock solution to (1.3) is admissible if the problem

$$\frac{d^2\hat{w}}{d\zeta^2} = -2s\frac{d\hat{w}(\zeta)}{d\zeta} - s^2(\hat{w}(\zeta) - w_1) - (p(\hat{w}(\zeta)) - p(w_1)) \tag{1.7a}$$

$$\hat{w}(-\infty) = w_1, \quad \hat{w}(+\infty) = w_2, \quad \hat{w}'(\pm\infty) = 0 \tag{1.7b}$$

has a solution. Some works related to the traveling wave problem (1.7) can be found in [7], [8], [10], [13], [20]–[26], [30], [31] and references cited therein.

Our approach of establishing the existence of solutions of the Riemann problem (1.3) with the initial value (1.4) is to construct the solutions as a $\epsilon \to 0+$ limit of solutions of

$$u_t + p(w)_x = \epsilon t u_{xx}, \tag{1.8a}$$

$$w_t - u_x = \epsilon t w_{xx}, \qquad x \in \mathbb{R}, t > 0, \tag{1.8b}$$

and (1.4). We shall call solutions of (1.3), (1.4) constructed in this way the solutions satisfying similarity viscosity criterion. This approach was pursued by many authors [2]–[6], [8]–[10], [16], [27], [28], [32]. We can see that $(u(x/t), w(x/t))$ is a solution of (1.8) and (1.4) if and only if it is a solution of

$$\epsilon u'' = -\xi u' + p(w)', \tag{1.9a}$$

$$\epsilon w'' = -\xi w' - u', \tag{1.9b}$$

$$(u(-\infty), w(-\infty)) = (u_-, w_-), \qquad (u(+\infty), w(+\infty)) = (u_+, w_+), \tag{1.9c}$$

where $\xi = x/t$ and " $'$ " is $d/d\xi$.

For this case, where w_- and w_+ are separated by the interval (α, β), Slemrod [27] proved the existence of (1.9). Fan [6] showed that there is a sequence $\{\epsilon_n\}$, $\epsilon_n \to 0+$ as $n \to \infty$, such that $(u_{\epsilon_n}(\xi), w_{\epsilon_n}(\xi))$, provided by Slemrod in [27], converges to a weak solution of the Riemann problem (1.3) as $n \to \infty$. The condition needed in [6] is, besides (1.3c),

$$|p(w)| \to \infty \quad \text{as } |w| \to \infty. \tag{1.10}$$

The solution constructed in [27] and [6] is also admissible by the traveling wave criterion (1.7) (cf. [8]).

The existence of the Riemann problem (1.3) in the case $w_\pm < \alpha$, $u_- < u_+$ (or $w_\pm > \beta$, $u_- > u_+$) remains, in general, unproved. For convenience, we shall call the Riemann problems with initial data lying in the same phase region the one-phase Riemann problem. An understanding of the Riemann problem in this case is necessary for the study of the interactions between phase boundaries.

In this paper, we shall prove, as one of our results, the existence of the Riemann for any initial data outside of the elliptic region. We state this result more precisely in the following theorem.

THEOREM 1.1. *Suppose $|p(w)| \to \infty$ as $|w| \to \infty$. Then for any Riemann initial data lying outside of the elliptic region, there is a solution of the Riemann problem* (1.3), (1.4) *satisfying the similarity viscosity criterion and possessing at most two phase boundaries. Furthermore, this solution is also admissible according to the traveling wave criterion* (1.7). *As a consequence, solutions of* (1.3), (1.4), *which are admissible by the traveling wave criterion, always exist.*

After we establish the existence of the Riemann problem, we proceed to study the structure of these solutions. Then we apply our results to study the outcome of the interactions between phase boundaries and between a phase boundary and an "ordinary" shock.

This paper is divided into six sections. In §2 we utilize the Leray–Shauder type fixed point theory to prove that (1.9) has a solution if some a priori estimates are given. In §3 we prove the estimates needed in §2. We also prove that the total variations of the solutions, $(u_\epsilon(\xi), w_\epsilon(\xi))$, of (1.9) constructed in §2 are independent of $\epsilon > 0$. Then, by the well-known Helly theorem, we can extract a sequence $\{\epsilon_n\}$ such that $\epsilon_n \to 0+$ as $n \to \infty$ and $(u_{\epsilon_n}(\xi), w_{\epsilon_n}(\xi))$ converges almost everywhere to a weak solution of (1.3), and hence the first part of Theorem 1.1 is proved. We state these results more precisely in §4. In §5 we prove the second part of Theorem 1.1. We also study the structure of solutions of the Riemann problem. In §6 we continue to study the structure of the solutions under the assumption that $p(w)$ is convex for $w \leq \alpha$ and concave for $w \geq \beta$, which makes the description of the structure of the solutions easier. After collecting information about the structures of solutions, we investigate the interactions among phase boundaries and "ordinary" shocks.

2. Existence of the connecting orbit $(u_\epsilon(\xi), w_\epsilon(\xi))$. In this section, we shall prove that the boundary value problem (1.9) has a solution, provided some a priori estimates which we shall establish in the next section are given.

We consider, instead of (1.9), the following altered system:

$$\varepsilon u'' = -\xi u' + \mu p(w)', \tag{2.1a}$$

$$\varepsilon w'' = -\xi w' - \mu u', \tag{2.1b}$$

$$(u(\pm L),\ w(\pm L)) = (u_\pm,\ w_\pm), \tag{2.1c}$$

where $L > 1$, $0 \leq \mu \leq 1$.

First of all, we recall some preliminary results from [2] and [27]. The following lemma is from [2, Thm. 4.1] or [27, Lemma 2.1].

LEMMA 2.1. *Let $(u_\epsilon(\xi), w_\epsilon(\xi))$ be the solution of* (1.3). *Then one of the following holds on any subinterval (a, b) for which $p'(w_\epsilon(\xi)) < 0$:*

(1) *Both $u_\epsilon(\xi)$ and $w_\epsilon(\xi)$ are monotone on (a, b);*

(2) *One of the $u_\epsilon(\xi)$ and $w_\epsilon(\xi)$ is a strictly increasing (decreasing) function with no critical point on (a, b) while the other has at most one critical point that is necessarily a local maximum (minimum) point.*

Now, we rewrite Lemma 2.2 of [27], which describes the shape of a solution of (1.9) in the elliptic region $\{(u.w) \in \mathbb{R}^2 \ : \ w \in (\alpha, \beta)\}$.

LEMMA 2.2. *Let $(u(\xi), w(\xi))$ be a solution of* (2.1) *with $\mu > 0$. Then on any interval $(l_1, l_2) \subset (-L, L)$ for which $p'(w(\xi)) > 0$ the graph of $u(\xi)$ versus $w(\xi)$ is convex at points where $w'(\xi) > 0$ and concave at points where $w'(\xi) < 0$.*

By considering (2.1), the existence of the connecting orbit problem (1.9) can be proved, as shown in the following theorem.

THEOREM 2.3. *Suppose* $u_- < u_+$ *and* $w_\pm < \alpha$. *Then there is a solution of* (1.9) *satisfying*

$$(u(\xi_1), w(\xi_1)) \neq (u(\xi_2), w(\xi_2)) \quad \text{for any } \xi_1, \xi_2 \in (-\infty, +\infty), \xi_1 \neq \xi_2, \\ w(\xi_1) \geq \bar{w} := \min(w_-, w_+), \tag{2.2}$$

and that

(2.3) *there are at most two disjoint open intervals*(a, b) *such that*

$$w(\xi) \in (\bar{w}, \alpha) \quad \text{for } \xi \in (a, b), \quad \text{and} \tag{2.3a}$$

$$\text{either } w(a) = \bar{w}, \quad w(b) = \alpha \quad \text{or} \quad w(a) = \alpha, \quad w(b) = \bar{w}, \tag{2.3b}$$

provided that the possible solution of (2.1) *satisfying* (2.2) *and* (2.3) *is bounded in* $C^1([-L, +L])$,

$$\|(u, w)\|_{C^1([-L.L];\mathbb{R}^2)} < M,$$

for some $M > 0$ *independent of* $\mu \in [0, 1]$ *and* $L > 1$.

Proof. We rewrite (2.1) as

$$\varepsilon y''(\xi) = \mu f(y)' - \xi y'(\xi), \tag{2.4}$$

where

$$y(\xi) = \begin{pmatrix} u(\xi) \\ w(\xi) \end{pmatrix}, \qquad f(y(\xi)) = \begin{pmatrix} p(w) \\ -u(\xi) \end{pmatrix}.$$

A straightforward calculation shows

$$\begin{aligned} y(\xi) = & y(-L) + z(y) \int_{-L}^{\xi} \exp\left(\frac{-\zeta^2}{2\varepsilon}\right) d\tau + \frac{\mu}{\varepsilon} \int_{-L}^{\xi} f(y(\tau)) d\tau \\ & - \frac{\mu^2}{\varepsilon} \int_{-L}^{\xi} \int_{-L}^{\zeta} \tau f(y(\tau)) \exp\left(\frac{\tau^2 - \zeta^2}{2\varepsilon}\right) d\tau d\zeta, \end{aligned} \tag{2.5}$$

where

$$\begin{aligned} z(x) = & \frac{1}{\int_{-L}^{L} \exp(-\xi^2/2\varepsilon) d\xi} \left[y(+L) - y(-L) - \frac{\mu}{\varepsilon} \int_{-L}^{L} f(x(\tau)) d\tau \right. \\ & \left. + \frac{\mu}{\varepsilon^2} \int_{-L}^{L} \int_{-L}^{\zeta} \tau f(x(\tau)) \exp\left(\frac{\tau^2 - \zeta^2}{2\varepsilon}\right) d\tau d\zeta \right] \\ = & z_1(x) + \mu z_2(x). \end{aligned} \tag{2.6}$$

Choose $\eta \in (\bar{w}, \alpha)$. We are interested in those functions $(u(\xi), w(\xi)) \in C^1([-L, +L]; \mathbb{R}^2)$ satisfying

$$(u(\xi_1), w(\xi_1)) \neq (u(\xi_2), w(\xi_2)) \quad \text{for any } \xi_1, \xi_2 \in [-L, +L], \xi_1 \neq \xi_2, \\ \text{and } w(\xi_1) \geq \eta \tag{2.7}$$

and

$$\text{(2.8)}\qquad \begin{aligned}&\text{there are at most two disjoint open intervals } (a,b) \text{ such that}\\ &w(\xi)\in(\eta,\alpha)\quad\text{for } \xi\in(a,b),\quad\text{and}\\ &\text{either } w(a)=\eta,\quad w(b)=\alpha\quad\text{or}\quad w(a)=\alpha,\quad w(b)=\eta.\end{aligned}$$

Now, we consider the open subset in $C^1([-L,+L];\mathbb{R}^2)$:

$$\text{(2.9)}\qquad \begin{aligned}\Omega :=\bigl\{(u,w)\in C^1([-L,L];\mathbb{R}^2)\ :\ &\|(u,w)\|_{C^1([-L.L];\mathbb{R}^2)}<M+1,\\ &\text{and (2.7) and (2.8) are satisfied}\bigr\}.\end{aligned}$$

We define an integral operator

$$T:\overline{\Omega}\times[0,1]\to C^1([-L,L];\mathbb{R}^2)$$

by

$$\text{(2.10)}\qquad \begin{aligned}T(x,\mu)(\xi)=&y(-L)+z(x)\int_{-L}^{\xi}\exp\left(\frac{-\zeta^2}{2\varepsilon}\right)d\zeta+\frac{\mu}{\varepsilon}\int_{-L}^{\xi}f(x(\zeta))d\zeta\\ &-\frac{\mu}{\varepsilon^2}\int_{-L}^{\xi}\int_{-L}^{\zeta}\tau f(x(\tau))\exp\left(\frac{\tau^2-\zeta^2}{2\varepsilon}\right)d\tau d\zeta,\end{aligned}$$

where $z(x)$ is given by (2.8). It is clear that a fixed point of $T(x,\mu)$ is a solution of (2.1).

It is a matter of routine analysis to show that T maps $\overline{\Omega}\times[0,1]$ continuously into $C^1([-L,L];\mathbb{R}^2)$. Furthermore, we can verify, by taking $d/d\xi$ twice on (2.10), that

$$\varepsilon(T(x.\mu)(\xi))''=\mu f(x(\xi))'-\xi(T(x.\mu)(\xi))'.$$

This implies that T maps $\overline{\Omega}\times[0,1]$ into a bounded, with bound independent of μ, subset of $C^2([-L,L];\mathbb{R}^2)$. Thus T is a compact operator from $C^1([-L.L];\mathbb{R}^2)\times[0,1]$ into $C^1([-L,L];\mathbb{R}^2)$.

We recall the following fixed point theorem (see Mawhin [19, Thm. IV.1]).

PROPOSITION 2.4. *Let X be a real normed vector space and Ω a bounded open subset of X. Let $T:\overline{\Omega}\times[0,1]\to X$ be a compact operator. If*

(i) *$T(x,\mu)\neq x$ for $x\in\partial\Omega$, $\mu\in[0,1]$, and*

(ii) *the Leray-Shauder degree $D_I(T(x,0)-x,\Omega,)\neq 0$,*

then $T(x,1)=x$ has at least one solution in Ω.

To solve our problem, we take $X=C^1([-L,+L];\mathbb{R}^2)$. We can see that (ii) is satisfied. Indeed,

$$\text{(2.11)}\quad T(x,0)-x=\frac{y(L)-y(-L)}{\int_{-L}^{L}\exp(-\zeta^2/2\varepsilon)\,d\zeta}\int_{-L}^{\xi}\exp\left(-\frac{\zeta^2}{2\varepsilon}\right)d\zeta+y(-L)-x=:x_0-x,$$

where $x_0\in\Omega$, and hence the degree $D_I(T(x,0)-x,\Omega,)=1$.

Now, we preceed to verify (ii) of Proposition 2.4. We assume, for contradiction, that there is a fixed point of $T(x, \mu)$,

$$x = (u, w)(\xi) \in \partial\Omega. \tag{2.12}$$

Then one of the following cases must hold.

Case A. $\|(u(\xi), w(\xi))\|_{C^1([-L,L];\mathbb{R}^2)} = M + 1$.

This case is excluded by the inequality (2.4).

Case B. The condition (2.7) is violated.

In this case, there are $\xi_1, \xi_2 \in [-L, +L]$, $\xi_1 \neq \xi_2$, such that $(u(\xi_1), w(\xi_1)) = (u(\xi_2), w(\xi_2))$ and $w(\xi_1) \geq \eta$.

The curve $(u(\xi), w(\xi))$ in (u, w)-plane near $\xi = \xi_1$ and $\xi = \xi_2$ cannot go across each other[1], as shown in Fig. 2. Otherwise, the curve of $(u(\xi), w(\xi))$ plus a $C^1([-L, +L]; \mathbb{R}^2)$ perturbation still intersects itself and hence is not in Ω. Hence $x = (u(\xi), w(\xi))$ is not in $\partial\Omega$, which yields a contradiction.

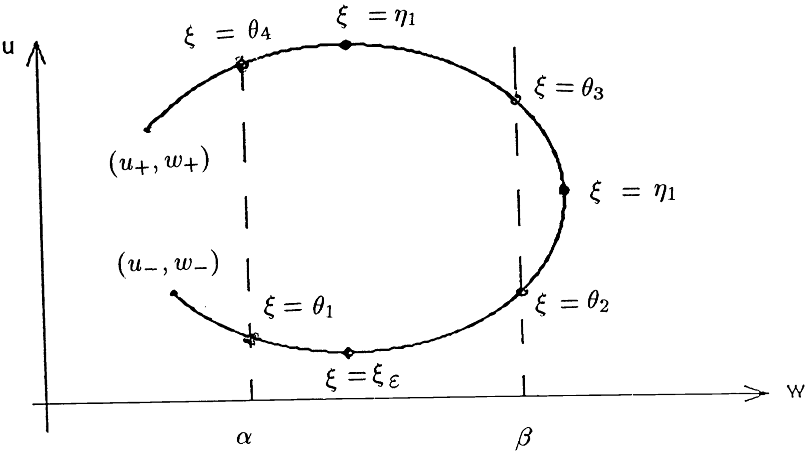


FIG. 2

From Lemma 2.1, we know that if $(u(\xi), w(\xi))$ stays inside the region $w \leq \alpha$, the curve $(u(\xi), w(\xi))$ cannot intersect itself. Thus, $w(\xi_3) > \alpha$ for some $\xi_3 \in [-L, +L]$.

We can further describe the curve $(u(\xi), w(\xi))$ in the (u, w)-plane as follows. There is an interval $[-L, \theta_1]$ such that $w(\xi) \leq \alpha$ and $w(\theta_1) = \alpha$, and by Lemma 2.1, $w'(\theta_1) > 0$. As ξ increases from θ_1, $(u(\xi), w(\xi))$ moves into the region $\alpha < w < \beta$. As long as $w'(\xi) > 0$ and $w(\xi) \in (\alpha, \beta)$, the curve $(u(\xi), w(\xi))$ in the (u, w)-plane

[1] We can make the above description rigorous by specifying a normal vector field $\mathbf{n}(\xi)$ along the curve $(u(\xi), w(\xi))$. Let y be the point $(u(\xi), w(\xi) + \mathbf{n}(\xi))$. Suppose the line connecting y and $(u(\xi), w(\xi))$, for ξ near ξ_1, intersects the portion of the curve $(u(\xi), w(\xi))$ in (u, w)-plane near $\xi = \xi_2$ at $(\bar{u}, \bar{w})$. Then we say the two portions of the curve $(u(\xi), w(\xi))$, near $\xi = \xi_1$ and ξ_2, respectively, do not go across each other if there is an interval $(\xi_1 - \theta_1, \xi_1 + \theta_2)$ such that

$$g(\xi) := ((\bar{u}, \bar{w}) - (u(\xi), w(\xi)))\mathbf{n}(\xi) \geq 0 \quad (\leq 0)$$

for $\xi \in (\xi_1 - \theta_1, \xi_1 + \theta_2)$ and

$$g(\xi_1 - \theta_1)g(\xi_1 + \theta_2) > 0.$$

The rest of the description of the curve can also be made rigorous in the same way.

is convex with respect to w. Let (θ_1, θ_2) be the largest interval such that $w'(\xi) > 0$ and $w(\xi) \in (\alpha, \beta)$. Then either $w(\theta_2) = \alpha$ or $w(\theta_2) \in (\alpha, \beta)$ and $w'(\theta_2) = 0$ hold. We assume, without loss of generality, that $w(\theta_2) = \beta$ and $w'(\theta_2) > 0$, since other cases are simpler. In view of Lemma 2.1, this interval is followed by another interval $[\theta_2, \theta_3]$ in which $w(\xi) \geq \beta$ and $u'(\xi) > 0$, while $w(\xi)$ has one and only one critical point which is a local maxima, and $w(\theta_3) = \beta$. Then there is the maximum interval $[\theta_3, \theta_4)$ in which $w(\xi) \in [\alpha, \beta]$, $w'(\xi) < 0$, and the curve $(u(\xi), w(\xi))$ in the (u, w)-plane is concave with respect to w. We see that at the right end of the interval, either $w(\theta_4) = \alpha$ or $w'(\theta_4) = 0$ must be true. Let us assume $w(\theta_4) = \alpha$ and $w'(\theta_4) < 0$, since the argument for this case covers the other cases and leads to the same conclusion. Thus, as ξ increases from θ_4, $(u(\xi), w(\xi))$ moves into the region $w < \alpha$.

From the shape of the curve $(u(\xi), w(\xi))$ in the (u, w)-plane for $\xi \in [-L, \theta_4]$, it is clear that

$$\xi_2 > \theta_4, \tag{2.13}$$

since otherwise, the curve will cross itself in the region $w \geq \alpha$.

Following $[\theta_3, \theta_4]$ is the interval (θ_4, θ_5) in which $w(\xi) < \alpha$. We, of course, pick the largest such interval, or more precisely,

$$\theta_5 := \sup\{\zeta \in [\theta_4, +L] \ : \ w(\xi) < \alpha \text{ for } \xi \in (\theta_4, \zeta)\}. \tag{2.14}$$

One of the following cases must occur.

Subcase B1. $\theta_5 = L$.

In this case, the derivative $w'(\xi) < 0$ if $w(\xi) \in [\alpha, \eta]$ for $\xi \in [\theta_4, \theta_5) = [\theta_4, +L)$ because if otherwise, $w(\xi)$ would have at least two extreme points in $(\theta_4, +L)$ which is impossible by Lemma 2.1. Furthermore, the point ξ_2 must satisfy $\xi_2 \in (\theta_4, L)$ and

$$w(\xi_2) \in (\alpha, \eta] \quad \text{and} \quad w'(\xi_2) < 0. \tag{2.15}$$

Hence $\xi_1 \in [-L, \theta_1)$ and

$$w'(\xi_1) > 0 \quad \text{and} \quad w(\xi) > w(\xi_1) = w(\xi_2) \quad \text{for } \xi \in (\xi_1, \xi_2). \tag{2.16}$$

Integrating (2.1b) over $[\xi_1, \xi_2]$ and using $(u(\xi_1), w(\xi_1)) = (u(\xi_2), w(\xi_2))$, (2.15), and (2.16), we obtain

$$0 < \int_{\xi_1}^{\xi_2} [w(\xi) - w(\xi_2)] d\xi = \varepsilon[w'(\xi_2) - w'(\xi_1)] < 0,$$

which is a contradiction. Thus, this case cannot happen.

Subcase B2. $\theta_5 < +L$.

For this case, $w(\theta_5) = \alpha$, and $w'(\theta_5) \geq 0$, $u(\xi)$ is decreasing over $[\theta_4 . \theta_5]$. $w(\xi)$ has one and only one critical point ζ_1 which is necessarily a local minima, and

$$u'(\xi) < 0 \quad \text{for } \xi \in (\theta_4, \theta_5). \tag{2.17a}$$

We claim that

$$w(\xi) > \alpha = w(\theta_5) \quad \text{for } \xi \in (\theta_5, \theta_5 + \delta) \tag{2.17b}$$

for some $\delta > 0$. Otherwise, α is a local maxima of $w(\xi)$, and hence

$$w'(\theta_5) = 0 \quad \text{and} \quad w''(\theta_5) \leq 0.$$

Thus, (2.1b) yields

$$\mu u'(\theta_5) = -\epsilon w''(\theta_5) \geq 0.$$

This inequality and (2.17a) imply that $u'(\theta_5) = 0$ and $w'(\theta_5) = 0$. Then, by the uniqueness of the initial value problem for (2.1), $(u(\xi), w(\xi)) \equiv (u(\theta_5), w(\theta_5))$, which violates the condition (2.1c). This contradiction proves (2.17b).

If

$$w(\zeta_1) < \eta, \tag{2.18}$$

then there are at least three disjoint subintervals (a_i, b_i), $i = 1, 2, 3$ lying in $[-L, \theta_1]$, $[\theta_4, \zeta_1]$ and $(\zeta_1, \theta_5]$, respectively, such that

$$\begin{aligned} &w(a_i) = \eta, \quad w(b_i) = \alpha, \quad i = 1, 3 \\ &w(a_2) = \alpha, \qquad w(b_2) = \eta \end{aligned} \tag{2.19}$$

It is clear that because of (2.18), $(u(\xi), w(\xi))$ plus some C^1 perturbations still have three disjoint subintervals (a_i, b_i), $i = 1, 2, 3$ satisfying (2.19). This simply says that $(u(\xi), w(\xi)) \notin \partial\Omega$, which contradicts (2.12), and hence

$$w(\zeta_1) \geq \eta.$$

By Lemmas 2.1 and 2.2, the portion of the curve $(u(\xi), w(\xi))$ for $\xi \in [\theta_5, +L]$ is confined in the domain $w \geq \eta$, and hence $w(+L) \neq w_+$, which is again a contradiction. Thus, Case B dose not occur.

Case C. The condition (2.3) fails to hold.

In fact, the arguments we used to handle the Subcase B2 can be applied to this case and yield a contradiction. This shows that Case C cannot happen either.

Summarizing our analysis for the above three cases, we find that if $(u(\xi), w(\xi)) \in \partial\Omega$, then $x = (u(\xi), w(\xi))$ cannot be a fixed point of $T(x, \mu)$ for $\mu \in [0, 1]$. Applying Proposition 2.4, we see that $T(x, 1)$ has a fixed point.

To prove the existence of solutions of (1.9), we need to pass to the limit $L \to \infty$. We follow Dafermos [2] and extend $(u(\xi), w(\xi))$ as follows:

$$(u(\xi; L), w(\xi; L)) = \begin{matrix} (u_+, w_+), & \xi > L, \\ (u_-, w_-), & \xi < -L. \end{matrix}$$

By the hypothesis (2.4), we see that $\{(u(\cdot; L), w(\cdot; L))\}$ is precompact in $C((-\infty, \infty); \mathbb{R}^2)$. So, there is a sequence $L_n \to \infty$ as $n \to \infty$ such that

$$(u(\xi; L_n), w(\rho; L_n)) \to (u(\xi, \infty), w(\xi, \infty))$$

uniformly as $n \to \infty$. By integrating (2.1a), (2.1b) with $\mu = 1$ twice from ξ_0, we can prove that the limit $(u(\xi, \infty),\ w(\xi, \infty))$ satisfies (1.9a), (1.9b). It remains to prove that $(u(\pm\infty, \infty),\ w(\pm\infty, \infty)) = (u_\pm,\ w_\pm)$. To this end, we manipulate (1.9a), (1.9b) to obtain

$$\frac{d}{d\xi}\left(\exp\left(\frac{\xi^2}{2\varepsilon}\right) y'(\xi)\right) = \frac{1}{\varepsilon}\left[f(y(\xi))' \exp\left(\frac{\xi^2}{2\varepsilon}\right)\right]$$

or

$$\exp\left(\frac{\xi^2}{2\varepsilon}\right) y'(\xi) = y'(0) + \frac{1}{\varepsilon}\int_0^\xi \nabla f(y) y'(\zeta) \exp\left(\frac{\zeta^2}{2\varepsilon}\right) d\zeta. \tag{2.20}$$

Applying (2.4) and Gronwell's inequality on (2.15), we obtain

$$\begin{aligned} |y'(\xi)| &\le |y'(0)| \exp\left(\frac{2R|\xi| - \xi^2}{2\varepsilon}\right) \\ &\le M \exp\left(\frac{2R|\xi| - \xi^2}{2\varepsilon}\right), \end{aligned} \tag{2.21}$$

where $R > 0$ depends at most on M, ν, and $\varepsilon > 0$. Inequality (2.21) holds for $y(\xi; L)$ also. Then

$$(u(\pm\infty, \infty), w(\pm\infty, \infty)) = (u_\pm, w_\pm)$$

follows from (2.21) easily. It remains to prove that the solution $(u(\xi, \infty), w(\xi, \infty))$ constructed above satisfies (2.2) and (2.3). Indeed, the same reasoning for Cases B and C implies that $(u(\xi, \infty), w(\xi, \infty))$ satisfies (2.8) and (2.9) also. Since $\eta \in (\bar{w}, \alpha)$ is chosen arbitrarily, (2.2) and (2.3) hold for $(u(\xi, \infty), w(\xi, \infty))$. □

COROLLARY 2.5. *Let* $(u(\xi), w(\xi))$ *be a solution of* (2.1) *or* (1.9) *satisfying* (2.2), (2.3).

(i) *The subset of* $[-L, +L]$

$$\{\ \xi \in [-L, +L] \ \ : \ \ w(\xi) \le \alpha\}$$

has at most two connected components. Furthermore, each component must have $-L$ *or* $+L$ *as one of its endpoints.*

(ii) *The set*

$$\{\ \xi \in [-L, +L] \ \ : \ \ w(\xi) \in (\alpha, \beta)\}$$

consists of at most two connected components.

(iii) *The set*

$$\{\ \xi \in [-L, +L] \ \ : \ \ w(\xi) \ge \beta\},$$

if nonempty, is an interval.

Proof. This is proved in our discussion in the proof of Theorem 2.3, Case B. □

The assumption (2.4) in above theorem can be replaced by a weaker one, as stated in the following theorem.

THEOREM 2.6. *The conclusion of Theorem* 2.3 *remains valid if* (2.4a) *is replaced by*

$$\sup_{-L \le \xi \le L} (|u(\xi)| + |w(\xi)|) \le M_1,$$

where M_1 *is independent of* $\mu \in [0, 1]$ *and* $L > 1$.

Proof. The proof is the same as that of Theorem 1.3 in [23]. □

Theorems 2.3 and 2.6 give the conditions under which (1.9) has a connecting orbit for $w_\pm < \alpha$ and $u_- < u_+$. Slemrod [27] proved the following theorem for the case $w_\pm < \alpha$ and $u_- < u_+$.

THEOREM 2.7. *Assume that $w_{\pm} < \alpha$ and $u_{-} < u_{+}$. Then, there is a solution of* (1.9) *satisfying*

$$w(\xi) \leq \alpha, \tag{2.22}$$

if every possible solution of (2.1) *satisfies*

$$||(u(\xi), w(\xi))||_{C([-L,+L];\mathbb{R}^2)} \leq C \tag{2.23}$$

for some constant C independent of $\mu \in [0,1]$ and $L > 1$.

3. A priori estimates. In this section, we shall prove the a priori estimates needed in Theorems 2.3 and 2.7 as well as some ϵ-independent estimates.

THEOREM 3.1. *Suppose $w_{\pm} < \alpha$ and $u_{-} < u_{+}$. Let $(u_\epsilon(\xi), w_\epsilon(\xi))$ be a solution of* (2.1) *with the properties* (2.2) *and* (2.3). *Then,*

$$||u_\epsilon(\xi)||_{C([-L,+L];\mathbb{R}^2)} \leq C, \tag{3.1}$$

where C is, throughout this section, a constant independent of $\epsilon > 0, \mu \in [0,1]$, and $1 < L \leq +\infty$.

Proof. When $\mu = 0$, our assertion can be easily verified. Thus, we assume $\mu > 0$ in the rest of the proof. We first prove $u_\epsilon(\xi) \geq C$. Let ξ_ϵ be a local minimum point of $u_\epsilon(\xi)$. Then either

$$w_\epsilon(\xi_\epsilon) \notin (\alpha, \beta), \qquad w'_\epsilon(\xi_\epsilon) < 0 \tag{3.2}$$

or

$$w_\epsilon(\xi_\epsilon) \in (\alpha, \beta), \qquad w'_\epsilon(\xi_\epsilon) > 0 \tag{3.3}$$

hold.

Case A. (3.2) holds.

In this case, by Lemma 2.1, $w(\xi_\epsilon) < \beta$. Otherwise, both $u_\epsilon(\xi)$ and $w_\epsilon(\xi)$ would have critical points in the set $\{\xi \in [-L,+L] : w_\epsilon(\xi) \geq \beta\}$, which is an interval by Corollary 2.5. Thus, $w(\xi_\epsilon) \leq \alpha$. Hence

$$\xi_\epsilon \in \{\xi \in [-L,+L] \ : \ w_\epsilon(\xi) \leq \alpha\} = [-L, \theta_1] \cup [\theta_4, +L], \tag{3.4}$$

where $\theta_1 \leq \theta_4$. If $\theta_1 < \theta_4$ and $\xi_\epsilon \in [-L, \theta_1]$, then $w'(\xi_\epsilon) < 0$ implies that $w_\epsilon(\xi)$ also has a critical point in $[-L, \theta_1]$, which is prohibited by Lemma 2.1. Thus, $\xi_\epsilon \in [\theta_4, L]$. Performing a calculation on (2.1), we obtain

$$\epsilon \frac{d}{d\xi}\left(\frac{du_\epsilon(\xi)}{dw_\epsilon(\xi)}\right) = \left(\frac{du_\epsilon(\xi)}{dw_\epsilon(\xi)} - \sqrt{-p'(w_\epsilon(\xi))}\right)\left(\frac{du_\epsilon(\xi)}{dw_\epsilon(\xi)} + \sqrt{-p'(w_\epsilon(\xi))}\right). \tag{3.5a}$$

This implies that, as ξ increases, $du_\epsilon(\xi)/dw_\epsilon(\xi)$ is decreasing if $|du_\epsilon(\xi)/dw_\epsilon(\xi)| \leq \sqrt{-p'(w_\epsilon(\xi))}$ and is increasing if $|du_\epsilon(\xi)/dw_\epsilon(\xi)| \geq \sqrt{-p'(w_\epsilon(\xi))}$. Thus the "initial" condition

$$\frac{du_\epsilon(\xi)}{dw_\epsilon(\xi)}\Big|_{\xi=\xi_\epsilon} = 0 \tag{3.5b}$$

leads to that for $\xi \in [\theta_4, +L]$,

$$\left|\frac{du_\epsilon(\xi)}{dw_\epsilon(\xi)}\right| \leq \max_{w_+ \geq w \geq \beta}\left(\sqrt{-p'(w)}\right), \tag{3.6}$$

and hence

$$u_\epsilon(\xi) \geq u_+ + (\alpha - w_-) \max_{w \in [w_+, \alpha]}\left(\sqrt{-p'(w)}\right). \tag{3.7}$$

Case B. (3.3) holds.

By Corollary 2.5, $[-L, L]$ can be divided as

$$[-L, L] = [-L, \theta_1] \cup (\theta_1, \theta_2) \cup [\theta_2, \theta_3] \cup (\theta_3, \theta_4) \cup [\theta_4, +L], \tag{3.8}$$

where, of course $\theta_1 \leq \theta_2 \leq \theta_3 \leq \theta_4$, and

$$\{\xi \in [-L, +L] \ : \ w_\epsilon(\xi) \leq \alpha\} = [-L, \theta_1] \cup [\theta_4, L], \tag{3.9a}$$

$$\{\xi \in [-L, +L] \ : \ w_\epsilon(\xi) \in (\alpha, \beta)\} = (\theta_1, \theta_2) \cup (\theta_3, \theta_4), \tag{3.9b}$$

$$\{\xi \in [-L, +L] \ : \ w_\epsilon(\xi) \geq \beta\} = [\theta_2, \theta_3]. \tag{3.9c}$$

It is clear that when (3.3) holds, $\xi_\epsilon \in (\theta_1, \theta_2) \neq \emptyset$ (cf. Fig. 2).

If $\xi_\epsilon \leq 0$, then the argument in the proof of Theorem 3.2 of [6] applies and yields $u_\epsilon(\xi) \geq C$. Thus, we assume

$$\xi_\epsilon > 0 \tag{3.10}$$

in the sequel of this proof.

Since ξ_ϵ is a local minimum point of $u(\xi)$, $u_\epsilon'(\xi) > 0$ for $\xi \in (\xi_\epsilon, \xi_\epsilon + \delta)$ for some $\delta > 0$. Then we can define

$$\eta_1 := \sup\{\zeta > \xi_\epsilon \ : \ u_\epsilon'(\xi) > 0 \text{ for } \xi \in (\xi_\epsilon, \zeta)\}. \tag{3.11}$$

Since $w_\epsilon(\xi_\epsilon) \leq \alpha$ and $w_\epsilon'(\xi_\epsilon) > 0$, there is a local maximum point η_2 of $w_\epsilon(\xi)$ with $\eta_2 > \xi_\epsilon$. We can further require that η_2 is the least of such points, i.e.,

$$\eta_2 := \sup\{\zeta > \xi_\epsilon \ : \ w_\epsilon'(\zeta) > 0\}. \tag{3.12}$$

Then, by Lemmas 2.1 and 2.2, $\eta_1 \notin (\xi_\epsilon, \eta_2)$, and hence (cf. Fig. 2)

$$\eta_1 > \eta_2 > \xi_\epsilon. \tag{3.13}$$

By integrating (2.1a) on (ξ_ϵ, ξ), where $\xi \in (\xi_\epsilon, \eta_2)$, we obtain

$$0 < \epsilon u_\epsilon'(\xi) = \int_{\xi_\epsilon}^{\xi} -\zeta u_\epsilon'(\zeta) d\zeta + \mu[\, p(w_\epsilon(\xi)) - p(w_\epsilon(\xi_\epsilon))\,].$$

It follows from (3.10) and (3.11) that $-\xi u'_{\epsilon_n}(\xi) < 0$ for $\xi \in (\xi_\epsilon, \eta_1)$. Thus, in view of (3.3), we have

$$\begin{aligned} 0 < \epsilon u'_\epsilon(\xi) &\leq \mu[p(w_\epsilon(\xi)) - p(w_\epsilon(\xi_\epsilon))] \\ &\leq \mu[p(w_\epsilon(\xi)) - p(\alpha)]. \end{aligned} \tag{3.14}$$

Therefore,

$$\alpha < w_\epsilon(\eta_2) \leq w_1. \tag{3.15}$$

Equation (3.13) also yields a useful inequality:

$$0 < \epsilon u'_\epsilon(\xi) \leq \mu(p(\beta) - p(\alpha)) \tag{3.16}$$

for $\xi \in [\xi_\epsilon, \eta_1]$.

Using (2.1), we can obtain

$$\frac{d^2 w_\epsilon}{du_\epsilon^2}(\xi) = \frac{-1}{\epsilon u'_\epsilon(\xi)}\left[1 + p'(w_\epsilon(\xi))\left(\frac{dw_\epsilon(\xi)}{du_\epsilon(\xi)}\right)^2\right]. \tag{3.17}$$

Hence, if

$$\left|\frac{dw_\epsilon(\xi)}{du_\epsilon(\xi)}\right| \leq \frac{1}{2\max_{w\in[w_0,w_1]}(\sqrt{|p'(w)|})}$$

and $\xi \in [\xi_\epsilon, \eta_1]$, then

$$\frac{d^2 w_{\epsilon_n}}{du_{\epsilon_n}^2}(\xi) \leq \frac{-\mu}{2\epsilon u'_\epsilon(\xi)} \leq -\frac{1}{2(p(\beta) - p(\alpha))}. \tag{3.18}$$

Thus, as ξ decreases from η_2 to ξ_ϵ, $dw_\epsilon(\xi)/du_\epsilon(\xi)$ will increase from zero, and eventually

$$\left.\frac{dw_\epsilon(\xi)}{du_\epsilon(\xi)}\right|_{\xi=\eta_3} = \frac{1}{2\max_{w\in[w_0,w_1]}(\sqrt{|p'(w)|})} \tag{3.19}$$

for some $\eta_3 \in (\eta_2, \xi_\epsilon)$ Let

$$\eta_4 := \sup\{\eta_3 \in (\xi_\epsilon, \eta_2) \ : \ \text{(3.19) is satisfied}\}. \tag{3.20}$$

Then,

$$\begin{aligned} \frac{1}{2\max_{w\in[w_0,w_1]}(\sqrt{|p'(w)|})} &= \left.\frac{dw_\epsilon(\xi)}{du_\epsilon(\xi)}\right|_{\xi=\eta_4} - \left.\frac{dw_\epsilon(\xi)}{du_\epsilon(\xi)}\right|_{\xi=\eta_2} \\ &= \int_{u_\epsilon(\eta_2}^{ue(\eta_4)} \frac{d^2 w_{\epsilon_n}}{du_{\epsilon_n}^2}(\xi) d(u_\epsilon(\xi)) \geq \frac{u_\epsilon(\eta_2) - u_\epsilon(\eta_4)}{2(p(\beta) - p(\alpha))} \end{aligned}$$

or

$$0 \leq u_\epsilon(\eta_2) - u_\epsilon(\eta_4) \leq \frac{p(\beta) - p(\alpha)}{\max_{w\in[w_0,w_1]}(\sqrt{|p'(w)|})}. \tag{3.21}$$

From (3.18), we also see that

$$\frac{dw_\epsilon(\xi)}{du_\epsilon(\xi)} \geq \frac{1}{2\max_{w\in[w_0,w_1]}(\sqrt{|p'(w)|})}$$

for $\xi \in (\xi_\epsilon, \eta_4)$. Thus,

$$\begin{aligned} 0 \leq & u_\epsilon(\eta_2) - u_\epsilon(\xi_\epsilon) = u_\epsilon(\eta_2) - u_\epsilon(\eta_4) + \int_{u_\epsilon(\xi_\epsilon)}^{u_\epsilon(\eta_4)} \frac{dw_\epsilon(\xi)}{du_\epsilon(\xi)} d(u_\epsilon) \\ \leq & \frac{p(\beta)-p(\alpha)}{\max_{w\in[w_0,w_1]}(\sqrt{|p'(w)|})} + 2 \max_{w\in[w_0,w_1]} \left(\sqrt{|p'(w)|}\right)(w_1 - w_0). \end{aligned} \tag{3.22}$$

Similarly, we can prove that

$$\begin{aligned} 0 \leq & u_\epsilon(\eta_1) - u_\epsilon(\eta_2) \\ \leq & \frac{p(\beta)-p(\alpha)}{\max_{w\in[w_0,w_1]}(\sqrt{|p'(w)|})} + 2 \max_{w\in[w_0,w_1]} \left(\sqrt{|p'(w)|}\right)(w_1 - w_0). \end{aligned} \tag{3.23}$$

Then we obtain

$$u_\epsilon(\xi_\epsilon) \geq u_\epsilon(\eta_1) - \frac{2(p(\beta)-p(\alpha))}{\max_{w\in[w_0,w_1]}(\sqrt{|p'(w)|})} - 4 \max_{w\in[w_0,w_1]} \left(\sqrt{|p'(w)|}\right)(w_1 - w_0). \tag{3.24}$$

If $u_\epsilon(\eta_1) \geq u_+$, then (3.24) shows that $u_\epsilon(\xi)$ is bounded from below uniformly in $\epsilon > 0$, $\mu \in [0,1]$, and $L > 1$. Now, we devote our attention to the case when

$$u_\epsilon(\eta_1) < u_+.$$

Then, $\eta_1 < L$ because $u_\epsilon(L) = u_+$. By the definition (3.11), of η_1, $u'_\epsilon(\eta_1) = 0$. Then, by Lemmas 2.1 and 2.2, η_1 has to be an extreme point for $u_\epsilon(\xi)$. Since $u'_\epsilon(\xi) > 0$ for $\xi \in (\xi_\epsilon, \eta_1)$, η_1 is a local maximum point. Lemmas 2.1 and 2.2 imply that either

$$w'_\epsilon(\eta_1) > 0 \quad \text{and} \quad w_\epsilon(\eta_1) \notin (\alpha, \beta) \tag{3.25}$$

or

$$w'_\epsilon(\eta_1) < 0 \quad \text{and} \quad w_\epsilon(\eta_1) \in (\alpha, \beta). \tag{3.26}$$

The case (3.25) cannot happen because it implies that $\eta_1 \in [-L, \theta_1]$, which violates the known fact that $\eta_1 > \xi_\epsilon \in [\theta_1, \theta_2)$. Then (3.26) and (3.25) imply that there is a local minimum point $\eta_5 > \eta_4$ of $u_\epsilon(\xi)$, which satisfies

$$u'_\epsilon(\eta_5) = 0 \quad \text{and} \quad w_\epsilon(\eta_5) \leq \alpha. \tag{3.27}$$

Then our argument for the Case A applies and gives us

$$|u_+ - u_\epsilon(\eta_5)| \leq (\alpha - w_+) \max_{w\in[w_+,\alpha]} \left(\sqrt{-p'(w)}\right). \tag{3.28}$$

Using (3.28) in (3.24), we obtain the desired result

(3.29)

$$\begin{aligned}u_\epsilon(\xi_\epsilon) \geq& u_\epsilon(\eta_1) - \frac{2(p(\beta)-p(\alpha))}{\max_{w\in[w_0,w_1]}\left(\sqrt{|p'(w)|}\right)} - 4\max_{w\in[w_0,w_1]}\left(\sqrt{|p'(w)|}\right)(w_1-w_0)\\ \geq& u_\epsilon(\eta_5) - \frac{2(p(\beta)-p(\alpha))}{\max_{w\in[w_0,w_1]}\left(\sqrt{|p'(w)|}\right)} - 4\max_{w\in[w_0,w_1]}\left(\sqrt{|p'(w)|}\right)(w_1-w_0)\\ \geq& u_+ - (\alpha - w_+)\max_{w\in[w_+,\alpha]}\left(\sqrt{-p'(w)}\right) - \frac{2(p(\beta)-p(\alpha))}{\max_{w\in[w_0,w_1]}\left(\sqrt{|p'(w)|}\right)}\\ &- 4\max_{w\in[w_0,w_1]}\left(\sqrt{|p'(w)|}\right)(w_1-w_0),\end{aligned}$$

which proves that $u_\epsilon(\xi)$ is bounded from below uniformly in $\epsilon > 0, \mu \in [0,1]$, and $L > 1$.

Similarly, we can also prove that $u_\epsilon(\xi)$ is bounded from above uniformly in $\epsilon > 0, \mu \in [0,1]$, and $L > 1$. □

In the remainder of this paper, we adopt the following notation:

(3.30a) $$u^* := \sup\{u_\epsilon(\xi) | \xi \in \mathbb{R}, \epsilon \in (0,1)\},$$

(3.30b) $$u_* := \inf\{u_\epsilon(\xi) | \xi \in \mathbb{R}, \epsilon \in (0,1)\}.$$

Once we established the a priori estimates for $u_\epsilon(\xi)$, we can proceed to prove the following results for $w_\epsilon(\xi)$ by using an argument similar to the one used in [2].

THEOREM 3.2. *Assume $w_\pm < \alpha$ and $u_- < u_+$. Let $(u_\varepsilon(\xi), w_\varepsilon(\xi))$ be a solution of* (2.1) *satisfying* (2.2) *and* (2.3). *Then*

(i) $||w_\varepsilon(\xi)||_{C([-l,+l];\mathbb{R}^2)} \leq C(\varepsilon)$ *where $C(\varepsilon)$ is independent of $\mu \in [0,1], L > 1$;*

(ii) *if $|p(w)| \to \infty$, as $|w| \to \infty$, and if $\mu = 1$, then $||w_\varepsilon(\xi)||_{C([-l,+l];\mathbb{R}^2)} \leq C$, where C is independent of $L > 1$ and $\varepsilon > 0$.*

Proof. We only prove that $w(\xi)$ is bounded from above uniformly. The other part of the proof is similar and is omitted.

(i) Without loss of generality, we assume $w_\varepsilon(\xi)$ has a local maximum point τ_ε. Without loss of generality, we further assume that

(3.31) $$\tau_\varepsilon \leq 0.$$

By Lemmas 2.1 and 2.2,

(3.32) $$u'_\epsilon(\tau_\varepsilon) > 0.$$

We define

(3.33) $$\eta := \inf\{\ \xi < \tau_\varepsilon\ :\ w'_\epsilon(\xi) > 0\}.$$

It is clear that $w'_\epsilon(\eta) \geq 0$. Integrating (2.1b), we obtain

(3.34) $$0 \geq -\epsilon w'_\epsilon(\eta) = -\int_\eta^{\tau_\epsilon} \xi w'_\epsilon(\xi)d\xi + \mu(u_\epsilon(\eta) - u_\epsilon(\tau_\varepsilon)).$$

By the definition (3.33), we see that $\xi w_\epsilon'(\xi) \leq 0$ on (η, τ_ε), and hence

$$0 \geq \int_\eta^{\tau_\varepsilon} \xi w_\epsilon'(\xi) d\xi \geq \mu(u_\epsilon(\eta) - u_\epsilon(\tau_\varepsilon)) \geq u_* - u^*. \tag{3.35}$$

If $\theta \leq \min(-1, \tau_n)$, then

$$\int_\eta^\theta \xi w_\epsilon'(\xi) d\xi \leq -\int_\eta^\theta w_\epsilon'(\xi) d\xi = w_\epsilon(\eta) - w_\epsilon(\theta).$$

From the definition (3.33), we know that either $\eta = -L$ or η is a local minimum point of $u_\epsilon(\xi)$. In view of Lemmas 2.1 and 2.2, $w_\epsilon(\eta) \in [\min(w_-, w_+), \beta]$. Then the above inequality yields

$$w_\epsilon(\theta) \leq -\int_\eta^\theta \xi w_\epsilon'(\xi) d\xi + w_\epsilon(\eta) \leq u^* - u_* + \beta. \tag{3.36}$$

In other words, $w_\epsilon(\xi)$ is bounded from above uniformly in $\epsilon > 0, \mu \in [0,1]$, and $L > 1$ if $\xi \leq \min(-1, \tau_\varepsilon)$. For $\xi \in (-1, \tau_\varepsilon]$, we have, from (2.1b), that

$$0 \leq -\epsilon w_\epsilon'(\xi) = \int_\xi^{\tau_\epsilon} -\zeta w_\epsilon'(\zeta) d\zeta + \mu(u_\epsilon(\xi) - u_\epsilon(\tau_\varepsilon)) \geq u_* - u^*.$$

This implies that

$$w_\epsilon(\tau_\varepsilon) \leq w_\epsilon(-1) + C_1(\epsilon) \leq u^* - u_* + \beta + C_1(\epsilon).$$

Thus, the statement (i) is proved.

(ii) It remains to consider the case when $\tau_\varepsilon \in (-1, 0]$ and $\mu = 1$. For each ϵ, we can choose $\zeta \in (-2, -1)$ such that $u_\epsilon(\zeta) \leq u^* - u_*$. By integrating (2.1a), with $\mu = 1$, on $[\theta, \tau_\varepsilon]$, we obtain

$$\begin{aligned} p(w_\epsilon(\tau_\varepsilon)) &= \epsilon u_\epsilon'(\tau_\varepsilon) - \epsilon u_\epsilon'(\theta) + p(w_\epsilon(\theta)) - \int_\theta^{\tau_\epsilon} \xi u_\epsilon'(\xi) d\xi \\ &\leq -\epsilon u_\epsilon'(\theta) + p(w_\epsilon(\theta)) - \int_\theta^{\tau_\epsilon} \xi u_\epsilon'(\xi) d\xi. \end{aligned} \tag{3.37}$$

Every term on the right-hand side of (3.33) is bounded uniformly in $\epsilon > 0$ and $L > 1$. Thus, by virtue of the assumption on p in the theorem, $w_\epsilon(\tau_\varepsilon)$ are bounded from below uniformly in $\epsilon > 0$ and $L > 1$. □

THEOREM 3.3. *Assume $w_\pm < \alpha$ and $u_- > u_+$. Let $(u_\varepsilon(\xi), w_\varepsilon(\xi))$ be a possible solution of* (2.1) *satisfying $w_\epsilon(\xi) \leq \alpha$. Then*

(i) $||u_\epsilon(\xi)||_{C([-L,+L];\mathbb{R}^2)} \leq C$, *where C is a constant independent of $L > 1$, $\mu \in [0,1]$ and $\varepsilon > 0$;*

(ii) $||w_\epsilon(\xi)||_{C([-L,+L];\mathbb{R}^2)} \leq C(\varepsilon)$, *where $C(\varepsilon)$ is independent of $\mu \in [0,1], L > 1$;*

(iii) *if $|p(w)| \to \infty$, as $|w| \to \infty$, and if $\mu = 1$, then $||w_\epsilon(\xi)||_{C([-L,+L];\mathbb{R}^2)} \leq C$, where C is independent of $L > 1$ and $\varepsilon > 0$.*

Proof. The proof is almost the same as that of Theorem 3.2. □

4. The existence of solutions of the one-phase Riemann problem. Combining Theorems 2.3, 2.4, 3.1, and 3.2, we obtain the following result.

THEOREM 4.1. (i) *Assume* $w_\pm < \alpha$ *and* $u_- < u_+$. *There is a solution* $(u_\epsilon(\xi), w_\epsilon(\xi))$ *of* (1.9) *satisfying*

$$(4.1)\quad \begin{aligned} &(u_\epsilon(\xi_1), w_\epsilon(\xi_1)) \neq (u_\epsilon(\xi_2), w_\epsilon(\xi_2)) \quad \textit{for any } \xi_1, \xi_2 \in (-\infty, +\infty), \xi_1 \neq \xi_2, \\ &\qquad w_\epsilon(\xi_1) \geq \bar{w} := +\min(w_-, w_+), \end{aligned}$$

and

(4.2) *there are at most two disjoint open intervals* (a, b) *such that*

$$(4.2a)\qquad w_\epsilon(\xi) \in (\bar{w}, \alpha),$$

and

$$(4.2b)\qquad \textit{either } w_\epsilon(a) = \bar{w}, \quad w_\epsilon(b) = \alpha \quad \textit{or} \quad w_\epsilon(a) = \alpha, \quad w_\epsilon(b) = \bar{w}.$$

(ii) *For the case* $w_\pm < \alpha$ *and* $u_- > u_+$, *there is a solution of* (1.9) *satisfying* $w_\epsilon(\xi) \leq \alpha$.

(iii) *There is a subsequence* $\{\varepsilon_n\}$, $\varepsilon_n \to 0+$ *as* $n \to \infty$, *such that* $(u_{\varepsilon_n}(\xi), w_{\varepsilon_n}(\xi))$ *given in* (i) *and* (ii) *converges almost everywhere to a weak solution* $(u(\xi), w(\xi))$ *of the Riemann problem* (1.3) *and* (1.4). *Furthermore, the solutions we constructed have at most two phase boundaries.*

Proof. Parts (i) and (ii) of Theorems 2.6, 3.1, 3.2, and 3.3 provide the a priori estimates needed by Theorem 2.3 and 2.7. Thus, parts (i) and (ii) are established.

(iii) From Corollary 2.5, we know that the solutions of (1.9) provided in (i), $u_\epsilon(\xi)$, and $w_\epsilon(\xi)$ are piecewise monotone. Thus, $(u_\varepsilon(\xi), w_\varepsilon(\xi))$ given in (i) has total variation bounded uniformly in $\varepsilon > 0$. Now the same arguments used in the proof of Theorem 3.2 of [1] or Theorem 4.1 of [27] prove the first part of our assertion. The second statement follows directly from (4.1) and (4.2). □

5. Structure of solutions of one-phase solutions. Before we proceed to study wave interactions in our system (1.3a), (1.3b), we have to have a clear picture of the structure of solutions of Riemann problems (1.3) we obtained in the last section.

DEFINITION 5.1. A solution of (1.3a), (1.3b) is said to be admissible according to viscosity-capillarity traveling wave criterion (or traveling wave criterion for short) if

(i) At each point ξ_0 of discontinuity of $(u(\xi), w(\xi))$, $(u(\xi_0-), w(\xi_0-))$ and $(u(\xi_0+), w(\xi_0+))$ exist, and

(ii) There are $v_k \in \mathbb{R}, k = 1, 2, \dots, n \in \mathbf{N}$, and $v_1 = w(\xi_0-), v_n = w(\xi_0+)$ such that the following boundary value problems have a solution:

$$(5.1a)\qquad \frac{d^2\hat{w}(\zeta)}{d\zeta^2} = -2\xi_0 \frac{d\hat{w}(\zeta)}{d\zeta} + p(\xi_0-) - p(\hat{w}(\zeta)) - \xi_0^2(\hat{w}(\zeta) - w(\xi_0-)),$$

$$(5.1b)\qquad \hat{w}(-\infty) = v_k, \quad \hat{w}(+\infty) = v_{k+1}, \quad \hat{w}'(\pm\infty) = 0.$$

THEOREM 5.1. *If $p(w)$ has the property that any straight line in (w,p)-plane intersects the graph of $p(w)$ at finite many points, then the solutions of* (1.3) *given by Theorem* 3.9, *which are admissible by the similarity viscosity criterion, are also admissible by the traveling wave criterion. Hence, solutions of* (1.3), *which are admissible by the traveling wave criterion, always exist.*

This theorem was first proved in [8] for the case where w_- and w_+ separated by the elliptic region (α,β). Although the above theorem is for any initial data with $w_\pm \notin (\alpha,\beta)$, the proof is basically the same as in [8]. Thus, we omit the proof.

Using the same technique developed in [3] and generalized in [6], we can prove the following theorem.

THEOREM 5.2. *Let $(u(\xi),w(\xi))$ be a solution of* (1.3). *Suppose ξ_0 is a point of discontinuity of $(u(\xi),w(\xi))$. Then one of the following holds*:

$$\lambda(w(\xi_0+)) \le \xi_0 \le \lambda(w(\xi_0+)), \tag{5.2a}$$

$$-\lambda(w(\xi_0+)) \le \xi_0 \le \lambda(w(\xi_0+)), \tag{5.2b}$$

$$-\lambda(w(\xi_0+)) \le \xi_0 \le -\lambda(w(\xi_0+)). \tag{5.2c}$$

THEOREM 5.3. *Let $(u(\xi),w(\xi))$ be a solution of* (1.3) *which satisfies the traveling wave criterion. Then $w(\xi)$ does not take value in (α,β), i.e., $w(\xi) \notin (\alpha,\beta)$ for almost all $\xi \in \mathbb{R}$.*

Proof. Assume, for contradiction, that $w(\xi_0-) \in (\alpha,\beta)$ for some $\xi_0 \in \mathbb{R}$. We claim that $w(\xi+) = w(\xi_0-)$ for $\xi \in (\xi_0-\delta,\xi_0)$ for some $\delta > 0$. Otherwise, one of the following two cases will occur.

Case (i). $w(\xi)$ is continuous on $(\xi_0-\delta,\xi_0)$ for some $\delta > 0$ and there is a sequence $\{\xi_n\} \subset (\xi_0-\delta,\xi_0)$ such that $\xi_n \to \xi_0-$ as $n \to \infty$ and $w(\xi_n+) \ne w(\xi_0-)$.

Case (ii). There is a sequence of points of discontinuity of $(u(\xi),w(\xi))$ such that $\xi_n \to \xi_0-$ as $n \to \infty$.

Case (ii) cannot occur because $w(\xi_n\pm) \in (\alpha,\beta)$ for large n and the Rankine–Hugoniot conditions at $\xi = \xi_n$ cannot hold.

We claim that Case (i) is also impossible. Indeed, we can integrate (1.9) over (ξ_n,ξ_0) to get

$$\xi_0 \frac{\Delta_n u}{\Delta_n w} = p'(\theta) - \frac{1}{\Delta_n w}\int_{\xi_n+}^{\xi_0-} (u(\zeta) - u(\xi_n+))d\zeta, \tag{5.3a}$$

$$\frac{\Delta_n u}{\Delta_n w} = -\xi_0 - \frac{1}{\Delta_n w}\int_{\xi_n+}^{\xi_0-} [w(\xi_n+) - w(\zeta)]d\zeta, \tag{5.3b}$$

where $\Delta_n w := w(\xi_0-) - w(\xi_n+) > 0$, $\Delta_n u := u(\xi_0-) - u(\xi_n+)$, and

$$\theta \in (w(\xi_n+), w(\xi_0-)).$$

It follows that

$$\xi_0 \lim_{n\to\infty} \frac{\Delta_n u}{\Delta_n w} = p'(w(\xi_0-)),$$
$$\lim_{n\to\infty} \frac{\Delta_n u}{\Delta_n w} = -\xi_0.$$

Then we arrive at the contradiction

$$\left(\lim_{n\to\infty}\frac{\Delta_n u}{\Delta_n w}\right)^2 = -p'(w(\xi_0-)) < 0.$$

It follows that there exist ξ_1 and ξ_2, which are points of discontinuity of $(u(\xi), w(\xi))$ such that $\xi_1 < \xi_0 < \xi_2$ and

$$w(\xi) = w(\xi_0-) \in (\alpha, \beta) \quad \text{for } \xi \in (\xi_1, \xi_2).$$

Therefore, according to the traveling wave criterion, both boundary value problems

$$\text{(5.4a)} \qquad \frac{d^2\hat{w}}{d\zeta^2} = -2\xi_1\frac{d\hat{w}(\zeta)}{d\zeta} - \xi_1^2(\hat{w}(\zeta) - w(\xi_1+)) - (p(\hat{w}(\zeta)) - p(w(\xi_1+))),$$

$$\text{(5.4b)} \qquad \hat{w}(-\infty) = w(\xi_1-), \quad \hat{w}(+\infty) = w(\xi_1+),$$

$$\text{(5.4c)} \qquad \hat{w}'(\pm\infty) = 0,$$

and

$$\text{(5.5a)} \qquad \frac{d^2\hat{w}}{d\zeta^2} = -2\xi_2\frac{d\hat{w}(\zeta)}{d\zeta} - \xi_2^2(\hat{w}(\zeta) - w(\xi_2-)) - (p(\hat{w}(\zeta)) - p(w(\xi_2-))),$$

$$\text{(5.5b)} \qquad \hat{w}(-\infty) = w(\xi_2-), \quad \hat{w}(+\infty) = w(\xi_2+),$$

$$\text{(5.5c)} \qquad \hat{w}'(\pm\infty) = 0$$

have solutions. A straightforward calculation shows that the eigenvalue for the linearized (near $\hat{w}(\zeta) = w(\xi_1+)$) problem of (5.4) is $\lambda = -\xi_1 \pm \sqrt{-p'(w(\xi_1+))}$. It is clear that $w(\xi_1+) = w(\xi_2-)$ is a node of (5.4). Thus, in order for (5.4) to have a solution, it is necessary that $\xi_1 \geq 0$. On the other hand, however, the same analysis shows that the solvability of (5.5) implies $0 \geq \xi_2 > \xi_1 \geq 0$. This contradiction proves our assertion.

6. Wave interactions in systems of mixed type. In this section, we study wave interactions involving phase boundaries. The study of interactions between waves, for instance phase boundaries and ordinary shocks, can be reduced to a study of, just as in the case of hyperbolic systems, Riemann problems (1.3). We shall provide a mechanism to quickly determine, at least qualitatively, the outcome of wave interactions in some situations.

For simplicity, we assume that

$$\text{(6.1)} \qquad p''(w)(w - \eta) < 0$$

for some $\eta \in (\alpha, \beta)$.

In the sequel, when we say solutions of (1.3), we mean weak solutions of (1.3) satisfying the traveling wave criterion, unless stated differently.

We first study the Riemann problem of (1.3) with initial value

$$(u(x,0),w(x,0)) = \begin{cases} (u_-,w_-) & \text{for } x<0, \\ (u_+,w_+) & \text{for } x>0\ , \end{cases} \tag{6.2}$$

where

$$w_- < \alpha < \beta < w_+. \tag{6.3}$$

We shall call the Riemann problem with (6.3) the two-phase Riemann problem.

For two-phase Riemann problem (1.3), (6.2), and (6.3), we know, from §2 of [7], that the unique centered solution $(u(\xi),w(\xi))$ of the Riemann (1.3), (6.2), and (6.3) consists of a shock $\xi = s_2$ such that $w_1 := w(s_2-) \le \alpha < \beta \le w(s_2+) =: w_2$, and two constant states (u_-,w_-) and (u_+,w_+). (u_-,w_-) are joined to $(u(s_2-),w(s_2-))$ by either a backward shock $\xi = s_1 < 0$ if $w(s_2-) < w_-$ or a backward rarefaction wave if $w(s_2-) > w_-$. (u_+,w_+) is connected to $(u(s_2+),w(s_2+))$ by either a forward shock $\xi = s_3 > 0$ if $w(s_2+) > w_+$ or a forward rarefaction wave if $w(s_2+) < w_+$. Thus, we can denote a solution of (1.3), for simplicity, by $\{w_1,w_2,s_2\}$. All the shocks mentioned in this paragraph, except possibly the phase boundaries, satisfy the Lax's shock admissibility criterion (cf. [17], [18]).

The following lemma is Lemma 4.1 from [7].

LEMMA 6.1. (1.3) *has a solution* $\{w_1,w_2,s_2\}$ *if and only if the following conditions hold:*

$$F(w_1,w_2,s_2) = u_+ - u_-, \tag{6.4}$$

where $w_1 \to w_2$ *is a connection with speed* s_2 *and* $w_1 \le \alpha < \beta \le w_2$*,*

$$\begin{aligned} F(w_1,w_2,s_2) :=& - s_1(w_1 - w_-)H(w_- - w_1) + H(w_1 - w_-+) \int_{w_-}^{w_1} \lambda(w)dw \\ & - s_2(w_2 - w_1) - H(w_2 - w_+)s_3(w_+ - w_2) \\ & + H(w_+ - w_2+) \int_{w_+}^{w_2} \lambda(w)dw, \end{aligned} \tag{6.5}$$

where $H(w)$ *is the heaviside function and*

$$s_k^2 := -\frac{p(w_k) - p(w_{k-1})}{w_k - w_{k-1}}, \quad k = 1,2,3, \quad w_0 := w_-, w_3 := w_+; \tag{6.6}$$

$$s_1 < 0, \quad s_3 > 0, \quad s_1 < s_2 < s_3, \tag{6.7a}$$

$$-\lambda(w_1) < s_1 < -\lambda(w_-), \quad \lambda(w_+) < s_3 < \lambda(w_2), \tag{6.7b}$$

$$-\lambda(w_2) < s_2 < \lambda(w_1), \tag{6.7c}$$

$$s_2 \le \lambda(w_2) \quad \textit{if } w_2 \ne w_+, \tag{6.7d}$$

$$s_2 \ge -\lambda(w_1) \quad \textit{if } w_1 \ne w_-. \tag{6.7e}$$

LEMMA 6.2 (see [22]). *For each w_1, there is at most one w_2 such that $w_1 \rightarrow w_2$ is a saddle-saddle connection of speed $s \geq 0$. As a consequence, the speed s is also unique.*

LEMMA 6.3 (see [7]). *Let $w_1 \rightarrow w_2$ be a saddle-saddle connection of speed $s \geq 0$ and $\bar{w}_1 \rightarrow \bar{w}_2$ a connection of speed $\bar{s} \geq 0$. If $\bar{w}_1 < w_1 < \alpha$ or $\bar{w}_1 > w_1 > \beta$, then $\bar{s} > s \geq 0$.*

LEMMA 6.4. *For each fixed $w_\pm$ with $w_- < \alpha < \beta < w_+$, we define*

(6.8)
$$A(w_-, w_+) := \{\{w_1, w_2, s_2\} \ : \ \{w_1, w_2, s_2\} \text{ is a solution to the Riemann problem of (1.3), (6.2), and (6.3) for some } u_\pm \in \mathbb{R}\}.$$

Then

(i) *$A(w_-, w_+)$ is a graph of a function $(w_1(s), w_2(s), s)$. Furthermore, $w_1(s)$ is a strictly decreasing function if $s \geq 0$ and $w_2(s)$ is a strictly decreasing function of s if $s \leq 0$;*

(ii) *$F(w_1, w_2, s_2)$ is a strictly decreasing function of s_2.*

Proof. (i) Since $\{w_1, w_2, s_2\}$ is a solution of the Riemann problem of (1.3), by the definition of the notation of $\{w_1, w_2, s_2\}$, $w_1 \rightarrow w_2$ must be a connection of speed s_2. We suppose $s_2 \geq 0$. If $w_1 \rightarrow w_2$ is a saddle-saddle connection, then, by Lemmas 6.2 and 6.3, w_1 is uniquely determined by s_2. If $w_1 \rightarrow w_2$ is not a saddle-saddle connection, then it has to be a saddle-node connection, since $s_2 \geq 0$. In view of (6.7d), $w_2 = w_+$ in this case, and hence w_1 is uniquely determined by

$$s_2^2 = -\frac{p(w_+) - p(w_1)}{w_+ - w_1}.$$

Thus, to prove that w_1 is a function of s_2, it suffices to prove that for each $s_2 \geq 0$, there cannot be two solutions $\{w_1, w_2, s_2\}$, $\{\bar{w}_1, \bar{w}_2, s_2\} \in A(w_-, w_+)$, where $w_1 \rightarrow w_2$ and $\bar{w}_1 \rightarrow \bar{w}_2$ are connections of the same speed s_2 and one of them is a saddle-saddle connection while the other is a saddle-node connection. To this end, we, without loss of generality, assume $\bar{w}_1 < w_1 < \alpha$. Then, by Lemma 6.3, $w_1 \rightarrow w_2$ must be a saddle-node connection, inferring $w_2 = w_+$, and $\bar{w}_1 \rightarrow \bar{w}_2$ is a saddle-saddle connection. Inspecting the graph of $p(w)$, noticing (6.1) and

$$\lambda(w_1) > s > \lambda(w_2), \qquad s < \lambda(\bar{w}_j), \quad j = 1, 2,$$

we can see that

$$\bar{s}_3 < \bar{s}_2 = s_2,$$

where $\bar{s}_3^2 = -(p(\bar{w}_2) - p(w))/(\bar{w}_2 - w_+)$, which cannot be true by (6.7a). This contradiction shows that if $s_2 \geq 0$, the w_1 in $\{w_1, w_2, s_2\} \in A(w_-, w_+)$ is a function of s_2 and hence so is the w_2. Similarly, we can prove that if $s_2 < 0$, the w_2 in $\{w_1, w_2, s_2\} \in A(w_-, w_+)$ is a function of s_2 and hence so is the w_1. Thus, the first statement of (i) is proved. The rest of the assertion (i) follows from Lemma 6.3.

(ii) The spirit of the proof for (ii) is similar to the proof of Lemma 4.4 of [7]. Thus we omit the proof. □

The above lemmas hold for the Riemann problems with

$$w_- < \alpha < \beta < w_+. \tag{6.9}$$

For the Riemann problems with

$$w_- > \beta > \alpha > w_+, \tag{6.10}$$

we can use the transformation $(u, w) \to (-u, -w)$ to convert the Riemann problem with (6.10) to that of (6.9). This is because $(u(x,t), w(x,t))$ is an admissible solution of

$$\begin{aligned} u_t + p(w)_x &= 0, \\ w_t - u_x &= 0, \qquad x \in \mathbb{R}, t > 0, \\ (u(x,0), w(x,0)) &= \begin{cases} (u_-, w_-) & \text{for } x < 0, \\ (u_+, w_+) & \text{for } x > 0, \end{cases} \end{aligned}$$

if and only if $(\bar{u}(x,t), \bar{w}(x,t)) = (-u(x,t), -w(x,t))$ is an admissible solution of

$$\begin{aligned} \bar{u}_t + (-p(-\bar{w}))_x &= 0, \\ \bar{w}_t - \bar{u}_x &= 0, \qquad x \in \mathbb{R}, t > 0, \\ (\bar{u}(x,0), \bar{w}(x,0)) &= \begin{cases} (-u_-, -w_-) & \text{for } x < 0, \\ (-u_+, -w_+) & \text{for } x > 0. \end{cases} \end{aligned}$$

Since $-p(-w)$ satisfies all the assumptions we need on $p(w)$, our results for the case $w_- < \alpha < \beta < w_+$ also hold for the admissible solutions of (6.10) with $w_- > \beta > \alpha > w_+$.

We state the results on the Riemann problem for the case $w_- > \beta > \alpha > w_+$ as follows. The unique centered solution $(u(\xi), w(\xi))$ of the Riemann (1.3), (6.2), and (6.9) has a shock $\xi = s_2$ such that $w_1 := w(s_2+) \leq \alpha < \beta \leq w(s_2-) =: w_2$ and two constant states, (u_-, w_-) and (u_+, w_+). (u_-, w_-) is joined to $(u(s_2-), w(s_2-))$ by either a backward shock $\xi = s_1 < 0$ if $w(s_2-) < w_-$ or a backward rarefaction wave if $w(s_2-) > w_-$. (u_+, w_+) is connected to $(u(s_2+), w(s_2+))$ by either a forward shock $\xi = s_3 > 0$ if $w(s_2+) > w_+$ or a forward rarefaction wave if $w(s_2+) < w_+$. Thus, we can denote a solution of (1.3), for simplicity, by $\{w_1, w_2, s_2\}$.

We can also define the function $F(w_1, w_2, s_2)$ as in Lemma 6.1 for the case $w_- > \beta > \alpha > w_+$. We shall, however, use $G(w_1, w_2, s_2)$ to denote it, i.e., $G(w_1, w_2, s_2) = F(w_1, w_2, s_2)$, to make this case distinct to readers' eyes. It is clear that (1.3), (6.2), and (6.9) have an admissible solution $\{w_1, w_2, s_2\}$ if and only if $G(w_1, w_2, s_2) = u_+ - u_-$.

LEMMA 6.5. *For each fixed $w_\pm$ with $w_+ < \alpha < \beta < w_-$, we define*

$$\begin{aligned} B(w_-, w_+) := \{\{w_1, w_2, s_2\} \ &: \ \{w_1, w_2, s_2\} \textit{ is a solution to the} \\ &\textit{Riemann problem of } (1.3),\ (6.2), \textit{ and } (6.9) \textit{ for some } u_\pm \in \mathbb{R}\}. \end{aligned} \tag{6.11}$$

Then

(i) *$B(w_-, w_+)$ is a graph of a function $(w_1(s), w_2(s), s)$. Furthermore, $w_1(s)$ is a strictly increasing function if $s \geq 0$, and $w_2(s)$ is a strictly increasing function of s if $s \leq 0$;*

(ii) *$G(w_1, w_2, s_2) = \bar{G}(s_2)$ is a strictly increasing function of s_2.*

The results in Lemmas 6.1–6.4 are for the case w_- and w_+ are separated by the spinodal region (α, β). Now, we concentrate on the case $w_\pm < \alpha$.

LEMMA 6.6. *Assume $w_\pm < \alpha$. Let $(u(\xi), w(\xi))$ be a centered solution of the Riemann problem satisfying the traveling wave criterion. If $w(\xi) \leq \alpha$, then each point of discontinuity of $(u(\xi), w(\xi))$ satisfies $\lambda(w(\xi_0-)) > \xi_0 > \lambda(w(\xi_0+))$ or $-\lambda(w(\xi_0-)) > \xi_0 > -\lambda(w(\xi_0+))$.*

Proof. This assertion follows directly from the assumption (6.1) and Lemma 2.2 of [7]. □

Remark. The above proposition generally does not hold without assumption (6.1).

As an application of the above analysis, we consider the wave interaction problem of (1.3) with the initial data

$$(u(x,0), w(x,0)) = \begin{cases} (u_-, w_-) & \text{for } x < -1, \\ (u_2, w_2) & \text{for } 1 > x > -1, \\ (u_+, w_+) & \text{for } x > 1. \end{cases} \tag{6.12}$$

The outcome of the interaction of the two shock waves $\{(u_-, w_-), (u_2, w_2)\}$ and $\{(u_2, w_2), (u_+, w_+)\}$ is the solution of the Riemann problem of (1.3) with initial value

$$(u(x,0), w(x,0)) = \begin{cases} (u_-, w_-) & \text{for } x < 0, \\ (u_+, w_+) & \text{for } x > 0\ . \end{cases} \tag{6.13}$$

We are interested in the interactions between ordinary shocks and phase boundaries and that between phase boundaries. To list all the possibilities of the interactions would be lengthy and perhaps unnecessary. Here, we only study two examples.

Example 1. Let $\{(u_-, w_-,), (u_2, w_2)\}$ be a phase boundary such that $w_- \longrightarrow w_2$ is a saddle-saddle connection with speed $s_2 \geq 0$, i.e., $0 \leq s_2 < \lambda(w_-), \lambda(w_2)$. Let $\{u_1, w_1), (u_+, w_+)\}$ be an ordinary admissible shock solution of (1.3) with speed $s_3 < 0$, i.e., $\lambda(w_2) > s_3 > \lambda(w_+)$. We further assume $w_+ > \beta$.

In this case, the two shocks must interact. A calculation shows that

$$u_- + F(w_-, w_2, s_2; w_-, w_+) = u_- + \bar{F}(s_2) < u_+; \tag{6.14}$$

see Fig. 3. We know from Lemma 6.1 that $\{\bar{w}_1, \bar{w}_2, \bar{s}_2\}$ is a solution of the Riemann problem (1.3) and (6.12) if and only if

$$u_- + \bar{F}(\bar{s}_2) = u_+,$$

which infers that

$$\bar{F}(s_2) < \bar{F}(\bar{s}_2).$$

By Lemma 6.4,

$$\bar{s}_2 < s_2,$$

and hence, by the same lemma,

$$\alpha > \bar{w}_1 > w_-.$$

We depict the result of the interaction $\{\bar{w}_1, \bar{w}_2, \bar{s}_2\}$, where $\bar{w}_1 < w_-$ and $\bar{s}_2 > s_2$ in Fig. 4.

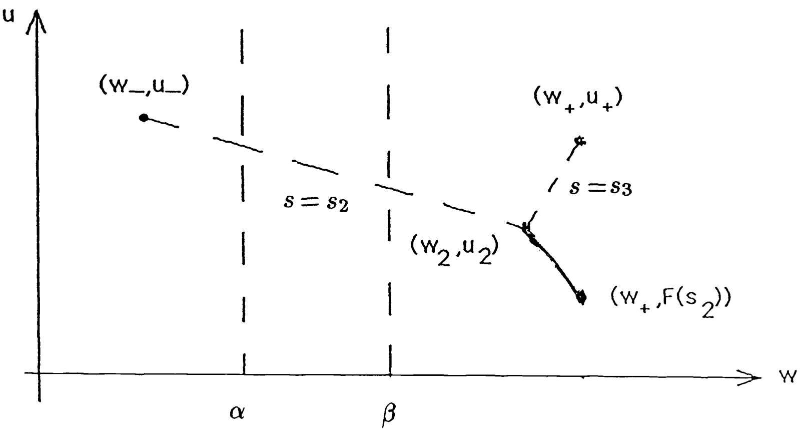


FIG. 3

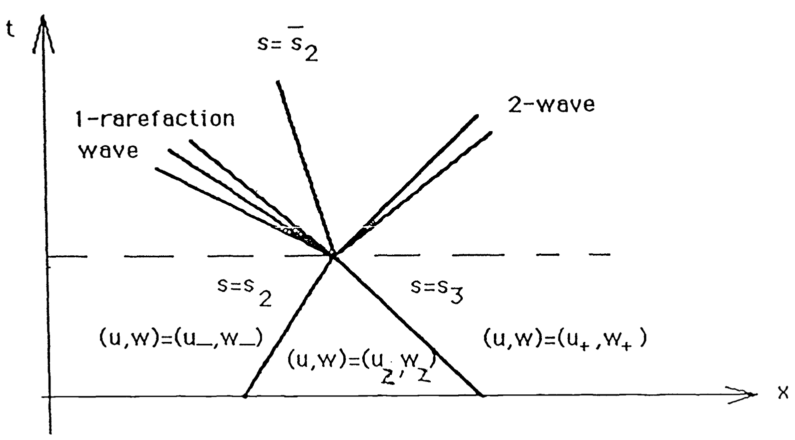


FIG. 4

Example 2. We consider the case when two phase boundaries interact. Let $\{(u_-, w_-,), (u_2, w_2)\}$ and $\{(u_2, w_2,), (u_+, w_+)\}$ be two phase boundaries with speed s_1 and s_2, respectively. We further assume that $w_\pm < \alpha$ and $w_2 > \beta$.

Above two phase boundaries interact if and only if $s_1 > s_2$. The outcome of the interaction is a solution of the Riemann problem (1.3), (6.11) with $w_\pm < \alpha$.

Case (i). $u_- \geq u_+$.

In this case, the solution of (1.3) and (6.11) satisfies $w(\xi) \leq \alpha$ and hence can be constructed easily by the classical method.

Case (ii). $u_- < u_+$.

Subcase (iia).

$$u_- < u_+ \leq u_- + \int_{w_-}^{\alpha} \sqrt{-p'(\eta)}d\eta + \int_{w_+}^{\alpha} \sqrt{-p'(\eta)}d\eta. \tag{6.15}$$

When (6.15) holds, there is no solution of (1.3), (6.13) satisfying $w(\xi) \leq \alpha$ and the Lax criterion (6.10). Then the solutions of (1.3), (6.11) must take values in the region $w > \beta$, i.e., $w(\xi) > \beta$ for some $\xi \in \mathbb{R}$.

Subcase (iib). $u_- < u_+$ but the second inequality of (6.15) is not satisfied.

In this case, the Riemann problem (1.3), (6.13) always has a solution which is confined in the region $w \leq \alpha$. Sometimes the Riemann problem also has another solution which satisfies $w(\xi) > \beta$ for some $\xi \in \mathbb{R}$ (cf. [22]). Although both of these solutions satisfy the traveling wave criterion, there is an example in [10] showing that, at least in that example, only the solution lying inside the region $w(\xi) \leq \alpha$ is admissible by the more "basic" vanishing viscosity criterion of the form

$$\begin{aligned} u_t + p(w)_x &= \epsilon u_{xx}, \\ w_t - u_x &= \epsilon w_{xx}, \end{aligned}$$

from which the traveling wave criterion is derived.

Acknowledgment. I would like to thank Professor M. Slemrod for his valuable suggestions and comments.

REFERENCES

[1] S. Antman, *Traveling waves in nonlinearly viscoelastic media and shock structure in elastic media*, Quart. Appl. Math., 46 (1988), pp. 77–93.

[2] C. M. Dafermos, *Solution of the Riemann problem for a class of hyperbolic conservation laws by the viscosity method*, Arch. Rational Mech. Anal., 52 (1973), pp. 1–9.

[3] ———, *Structure of solutions of the Riemann problem for hyperbolic systems of conservation laws*, Arch. Rational Mech. Anal., 53 (1974), pp. 203–217.

[4] C. M. Dafermos and R. J. DiPerna, *The Riemann problem for certain classes of hyperbolic systems of conservation laws*, J. Differential Equations, 20 (1976), pp. 90–114.

[5] H.-T. Fan, *The structure of solutions of the gas dynamics equations and the formation of the vacuum state*, Quart. Appl. Math., 49 (1991), pp. 29–48.

[6] ———, *A limiting "viscosity" approach to the Riemann problem for materials exhibiting changes of phase* (II), Arch. Rational Mech. Anal., 116 (1991), pp. 317–337.

[7] ———, *The uniqueness and stability of the solution of the Riemann problem for a system of conservation laws of mixed type*, Trans. Amer. Math. Soc., 333 (1992), pp. 913–938.

[8] ———, *The Existence, Uniqueness and Stability of the Solutions of the Riemann Problem of a System of Conservation Laws of Mixed Type*, Ph.D. thesis, University of Wisconsin-Madison, 1990.

[9] ———, *A vanishing viscosity approach on the dynamics of phase transitions in van der Waals fluids*, J. Differential Equations, to appear.

[10] ———, *Global versus local admissibility criteria for dynamics of phase transitions*, Proc. Roy. Soc. Edinburgh Sect. A, to appear.

[11] J. Glimm, *The interactions of nonlinear hyperbolic waves*, Comm. Pure Appl. Math., 41 (1988), pp. 569–590.

[12] ———, *Nonlinear and stochastic phenomena, the grand challenge for partial differential equations*, SIAM Rev., 33 (1991), pp. 626–643.

[13] R. Hagan and M. Slemrod, *The viscosity-capillarity admissibility criterion for shocks and phase transitions*, Arch. Rational Mech. Anal., 83 (1984), pp. 333–361.

[14] E. Isaacson, D. Marchesin, and B. Plohr, *Transitional waves for conservation laws*, SIAM J. Math. Anal., 21 (1990), pp. 837–866.

[15] E. Isaacson, D. Marchesin, C. F. Palmeira, and B. Plohr, *A global formalism for nonlinear waves in conservation laws*, preprint, 1991.

[16] A. S. Kalasnikov, *Construction of generalized solutions of quasi-linear equations of first order without convexity conditions as limits of solutions of parabolic equations with a small parameter*, Dokl. Akad. Nauk. SSSR 127 (1959), pp. 27–30. (In Russian.)

[17] P. D. Lax, *Hyperbolic systems of conservation laws*, Comm. Pure Appl. Math., 10 (1957), pp. 537- 566.

[18] T.-P.Liu, *The Riemann problem for general system of conservation laws*, J. Differential Equations, 18 (1975), pp. 218–234.

[19] M. Mawhin, *Topological degree methods in nonlinear boundary value problems*, CBMS Regional Conference Series in Mathematics No. 40, American Mathematical Society, Providence, RI, 1979.

[20] S. Schecter and M. Shearer, *Transversality for undercompressive shocks in Riemann problems*, in Viscous Profiles and Numerical Methods for Shock Waves, M. Shearer, ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1991.

[21] M. Shearer, *Admissibility criteria for shock wave solutions of a system of conservation laws of mixed type*, Proc. Roy. Soc. Edinburgh, 93 (1983), pp. 233–244.

[22] ———, *Nonuniqueness of admissible solutions of Riemann initial value problem for a system of conservation laws of mixed type*, Arch. Rational Mech. Anal., 93 (1986), pp. 45–59.

[23] ———, *Dynamic phase transitions in a van der Waals gas*, Quart. Appl. Math., 46 (1988), pp. 631–636.

[24] M. Slemrod, *Admissibility criterion for propagating phase boundaries in a van der Waals fluid*, Arch. Rational Mech. Anal., 81 (1983), pp. 301–315.

[25] ———, *Dynamic phase transitions in a van der Waals fluid*, J. Differential Equations, 52 (1984), pp. 1–23.

[26] ———, *Dynamics of first order phase transitions*, in Phase Transitions and Material Instabilities in Solids, M. Gurtin, ed., Academic Press, New York, 1984, pp. 163–203.

[27] ———, *A limiting "viscosity" approach to the Riemann problem for materials exhibiting change of phase*, Arch. Rational Mech. Anal. 105 (1989), pp. 327–365.

[28] M. Slemrod and A. Tzavaras, *A limiting viscosity approach for the Riemann problem in isentropic gas dynamics*, Indiana Univ. Math. J., 38 (1989), pp. 1047–1073.

[29] B. Temple, *Global existence of the Cauchy problem of a class of* 2×2 *nonstrictly hyperbolic conservation laws*, Adv. Appl. Math., 3 (1982), pp. 355–375.

[30] L. Truskinovskii, *Equilibrium phase interfaces*, Dokl. Akad. Nauk SSSR 265 (1982), pp. 306–310.

[31] ———, *Kinks versus shocks*, in the conference proceedings of Shock Induced Transitions and Phase Structures in General Media, Institute of Mathematics and its Applications, University of Minnesota, 1990, to appear.

[32] V.A.Tupciev, *The asymptotic behavior of the solutions of Cauchy problem for the equation* $\epsilon^2 t u_{xx} = u_t + [\phi(u)]_x$ *that degenerates for* $\xi = 0$ *into the problem of the decay of an arbitrary discontinuity for the case of a rarefaction wave,*, Z. Vycisl. Mat. Fiz., 12 (1972) pp. 770–775; English translation in USSR Comput. Math. Phys., 12.

SIAM J. MATH. ANAL.
Vol. 24, No. 4, pp. 866–891, July 1993

SOLITARY WAVES IN A TWO-LAYER FLUID WITH SURFACE TENSION*

S. M. SUN†‡ AND M. C. SHEN‡

Abstract. This paper deals with permanent gravity-capillary waves on the interface with surface tension in a two-layer, inviscid, and incompressible fluid between two horizontal, rigid boundaries. It is shown that, if the Bond number τ, a nondimensional surface tension coefficient, is greater than some critical value τ_0, and the Froude number F is less than, but near some critical value F_0, there exists a solitary wave solution which decays to zero at infinity. When τ is less than τ_0, solitary waves plus a small oscillation at infinity will appear, and the existence of such type of solutions will be investigated in a subsequent paper. Discussions about several critical cases, such as τ near τ_0 or a density ratio near some critical value, are also given.

Key words. solitary waves, two-layer fluids, surface tension

AMS subject classifications. 76B25, 76C10, 76B45, 35R35

1. Introduction.

This paper concerns the existence of an interface solitary wave between two immiscible, inviscid, and incompressible fluids of different but constant densities in the presence of surface tension. The upper and lower boundaries are assumed to be horizontal and rigid. We choose a coordinate system moving with the wave at a constant speed C so that in reference to this system the fluid motion is steady. Let h^-, h^+ be the depths at infinity, and ρ^-, ρ^+ be the densities of the lower and upper layers, respectively, T be the constant surface tension coefficient, and g be the constant gravitational acceleration. Define $h = h^+/h^-$, $\rho = \rho^+/\rho^- < 1$, the Froude number $F = C^2/((1-\rho)gh)$, and the Bond number $\tau = T/(h^-C^2\rho^-)$. Let us briefly review the existent results for a solitary wave on a layer of fluid over a bottom in the presence of surface tension. This case corresponds to $\rho_+ = 0$. Apparently Korteweg and de Vries [1] first derived a model equation under long wave approximation when F is near 1. The equation is now called the K-dV equation, and for $\tau = 0$ possesses a progressive wave solution, the so-called solitary wave, which decays to zero at infinity for $F > 1$ but near 1. The existence of such a solution to the exact equations was proved by Friedrichs and Hyers [2] among others half a century later. However, if the surface tension is taken into account in the K-dV equation, there is another critical value $\tau_0 = \frac{1}{3}$ of the Bond number τ. When τ is near $\frac{1}{3}$, the K-dV equation is no longer valid and a fifth-order model equation was derived by Hunter and Vanden-Broeck [3]. For $\tau > \frac{1}{3}$, $F < 1$ but near 1, the K-dV equation has a solitary wave of depression. The existence of this type of solution to the exact equations was verified numerically in [3] and proved by Amick and Kirchgässner [4] and Sachs [5]. For $\tau > \frac{1}{3}$ but near $\frac{1}{3}$ and $F < 1$ but near 1, the existence of a new solitary wave of depression was proved by Sun and Shen [6]. For $0 < \tau < \frac{1}{3}$, $F > 1$ but near 1, the solitary wave solution of the K-dV equation represents a wave of elevation. The behavior of a solution to the exact equations for this case is quite different. Hunter and Vanden-Broeck [3] also computed the solution numerically and found that it consists of a solitary wave of elevation plus

*Received by the editors February 20, 1991; accepted for publication (in revised form) August 10, 1992. This research was partly supported by National Science Foundation grant CMS-8903083.

†Present address: Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

‡Department of Mathematics, University of Wisconsin-Madison, Madison, Wisconsin 53706.

small oscillations at infinity, which we will call a generalized solitary wave. Hunter and Scheurle [7] studied a generalized solitary wave solution of the fifth-order equation for $\tau < \frac{1}{3}$ but near $\frac{1}{3}$ and $F > 1$ but near 1. The existence of a generalized solitary wave solution to the exact equations for $0 < \tau < \frac{1}{3}$, $F > 1$ but near 1 has been given recently by Beale [8] and Sun [9] on the basis of different methods.

For the case of a two-layer fluid without surface tension, the first systematic study is due to Peters and Stoker [10]. Amick and Turner [11] have recently used the center manifold technique to study possible interface waves, and computational results have been given by Turner and Vanden-Broeck [12]. An existence proof of internal solitary waves has also been given by Bona and Sachs [13], which gives many interesting properties of these solitary wave solutions as well. For a two-layer fluid in the presence of surface tension, the critical values of F and τ are, respectively, found as $F_0 = (1+\rho/h)^{-1}$ and $\tau_0 = (1+\rho h)/3$. A K-dV equation can be derived for F near F_0 and possesses an approximate solitary wave solution if $\rho \neq h^2$ but not near h^2, and $\tau \neq \tau_0$ but not near τ_0. The main objectives of this paper are to prove that a solitary wave solution of the K-dV equation indeed is an approximate solution to the exact equations if $\tau > \tau_0$ but not near τ_0, $\rho \neq h^2$ but not near h^2, and $F < F_0$ but near F_0, and to consider the critical cases when τ is near τ_0 or ρ is near h^2.

The main result of this paper is the following existence result: If $\nu = F^{-1} = 1 + (\rho/h) + \lambda_1\epsilon, \lambda_1 > 0$ and $\tau > \tau_0$, then for small $\epsilon > 0$ there exists an internal solitary wave solution in the form

$$f^+(x,\psi) = \psi + \epsilon(\psi - h)S(x) + \epsilon((\psi - h)\omega(x) + \theta^+(x,\psi)) \quad \text{for } \ 0 < \psi < h\,,$$

$$f^-(x,\psi) = \psi - h\epsilon(\psi + 1)S(x) + \epsilon(-h(\psi + 1)\omega(x) + \theta^-(x,\psi)) \quad \text{for } \ -1 < \psi < 0,$$

where f^+, f^- are the streamline functions for the upper and lower layers, respectively; ψ is the stream function

$$S(x) = \frac{\lambda_1 h}{h^2 - \rho}\operatorname{sech}^2\frac{(\lambda_1/(\tau - \tau_0))^{1/2}\epsilon^{1/2}x}{2},$$

and $\omega, \theta^{\pm}$ are error terms with $\|\omega(x)\|_{X^s} \leq K\epsilon$ and $\|\theta(x,\psi)\|_{B^s} \leq K\epsilon$ for $s \geq 4$. Here K is a constant independent of ϵ and X^s, B^s are Banach spaces to be defined later. The interface is given by $\eta = f^+(x,0) = f^-(x,0) = \epsilon(-h)S(x) + O(\epsilon^2)$. The outline of our existence proof of the solitary wave solution is as follows: The fluid domain with an unknown interface is first transformed to a horizontal strip with a fixed interface and boundaries. Then the exact equations are linearized about a given constant state, and a solvability condition appears for the solution of the corresponding inhomogeneous equations. Since the eigenvalue problem for the linearized equations has no periodic eigenfunction, a Banach space of functions decaying to zero at infinity can be defined. We can obtain an integro-differential equation from the solvability condition and several a priori estimates for a solution in the Banach space. The partial differential equations are now studied on the basis of a Hilbert space with an inner product specially constructed for a two-layer problem in the presence of surface tension. A solution of these equations is then expressed in terms of eigenfunctions of an eigenvalue problem associated with the linearized equations and a priori estimates for this solution in the Banach space are obtained. By these estimates and contraction mapping theorem, the existence of a solitary wave decaying to zero at infinity is proved. Our approach is based upon several ideas due to Friedrichs and Hyers [2], Beale [8], and Ter-Krikorov [14]. For the case $\tau < \tau_0$, the eigenvalue problem for the linearized

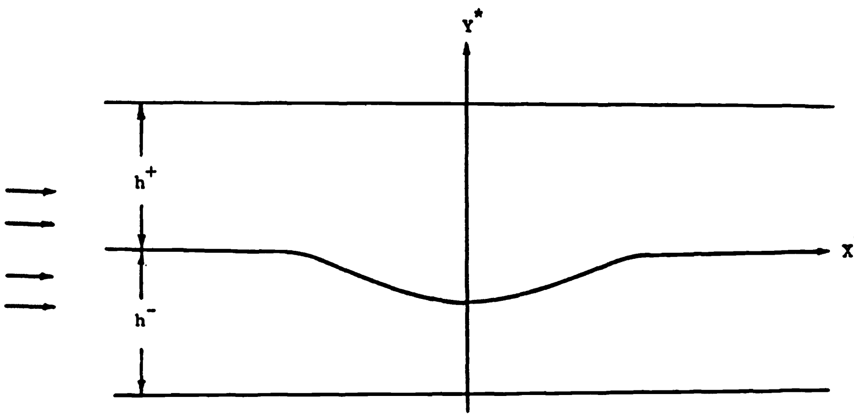


FIG. 1. *Configuration of a solitary wave.*

equations possesses a positive eigenvalue, and the corresponding eigenfunction is a periodic function. The main difficulty is to isolate the oscillatory part of a solution to the exact equations. We shall defer the existence proof of a solution for this case in a subsequent paper. Several critical cases also warrant consideration here. If τ is near τ_0 or ρ is near h^2, the present asymptotic theory is no longer valid and new asymptotic schemes have to be developed. We shall derive different approximate equations for the interface waves and obtain new solitary wave solutions to these equations. The justification of the formal derivation of a solitary wave solution for ρ near h^2 and $\tau > \tau_0$ but not near τ_0 is also presented. The justifications for the other cases will be deferred to subsequent investigation.

In §2, we formulate the problem in terms of nondimensional equations in a horizontal strip with a fixed interface. A formal derivation of the stationary K-dV equation and its solitary wave solution is presented in §3. An existence theorem of the approximate solitary wave solution to the exact equations is proved in §4. In §5, we discuss several critical cases for which the present asymptotic method fails, and we formally derive new model equations and their solitary wave solutions. In §6 we justify the formal method for ρ near h^2 and $\tau > \tau_0$ but not near τ_0. In the Appendix it is shown that there exists no positive eigenvalue of the linearized equations for $\tau > \tau_0$.

2. Formulation. We consider the irrotational flow of two immiscible, inviscid, and incompressible fluids of different but constant densities with surface tension at the interface and bounded by two horizontal rigid boundaries. In reference to the coordinate system moving at a speed C, the fluid flow becomes steady (Fig. 1), and the governing equations in terms of the stream function ψ^* are

$$\nabla^2\psi^{*-} = 0, \qquad -h^- < y^* < \eta^*(x^*), \tag{1}$$

$$\nabla^2\psi^{*+} = 0, \qquad \eta^*(x^*) < y^* < h^+, \tag{2}$$

where the velocity $\overline{v} = (u, v) = (\psi^*_{y^*}, -\psi^*_{x^*})$, and $y^* = \eta^*(x^*)$ is the interface. The boundary conditions are the following: At $y^* = \eta^*(x^*)$,

$$\eta^*_{x^*}\psi^{*\pm}_{y^*} + \psi^{*\pm}_{x^*} = 0, \tag{3}$$

$$\begin{aligned}(\rho^- - \rho^+)g\eta^* + \frac{\rho^-}{2}(\nabla\psi^{*-})^2 - \frac{\rho^+}{2}(\nabla\psi^{*+})^2 \\ - T\eta^*_{x^*x^*}(1 + (\eta^*_{x^*})^2)^{-3/2} = D;\end{aligned} \tag{4}$$

at $y^* = h^+$,

$$\psi^{*+}_{x^*} = 0; \tag{5}$$

at $y^* = -h^-$,

$$\psi^{*-}_{x^*} = 0. \tag{6}$$

Here D is the Bernoulli constant on the interface. Now we use x^*, ψ^* as independent variables and the so-called streamline function f^* as the dependent variable, where $\psi^*(x^*, f^*) =$ constant defines a streamline. It is obtained that

$$f^{*\pm}_{\psi^*} = \frac{1}{\psi^{*\pm}_{y^*}}, \qquad f^{*\pm}_{x^*} = -\frac{\psi^{*\pm}_{x^*}}{\psi^{*\pm}_{y^*}}.$$

By using the above relations, (1)–(6) become

$$(1 + (f^*_{x^*})^2) f^*_{\psi^*\psi^*} + (f^*_{\psi^*})^2 f^*_{x^*x^*} - 2 f^*_{\psi^*} f^*_{x^*} f^*_{x^*\psi^*} = 0 \tag{7}$$

for $0 < \psi^* < \psi_2$ and $-\psi_1 < \psi^* < 0$, where ψ_1 and ψ_2 are two constants and

$$f^* = \begin{cases} f^{*+}, & 0 < \psi^* < \psi_2, \\ f^{*-}, & -\psi_1 < \psi^* < 0. \end{cases}$$

If on $\psi^* = 0$, f^* goes to zero as $|x^*| \to \infty$, then $\psi_1 = Ch^-$ and $\psi_2 = Ch^+$. At $\psi^* = 0$,

$$f^{*+} = f^{*-}, \tag{8}$$

$$\begin{aligned} (\rho^- - \rho^+) g f^* - T f^*_{x^*x^*} (1 + (f^*_{x^*})^2)^{-3/2} + \frac{\rho^-}{2} \frac{1 + (f^{*-}_{x^*})^2}{(f^{*-}_{\psi^*})^2} \\ - \frac{\rho^+}{2} \frac{1 + (f^{*+}_{x^*})^2}{(f^{*+}_{\psi^*})^2} = D; \end{aligned} \tag{9}$$

at $\psi^* = -\psi_1$,

$$f^{*-} = -h^-; \tag{10}$$

at $\psi^* = \psi_2$,

$$f^{*+} = h^+. \tag{11}$$

To nondimensionalize (7)–(11), we measure ψ^* in units of ψ_1 so that $\psi = -1$ at bottom. Define

$$x = \frac{x^*}{h^-}, \quad \psi = \frac{\psi^*}{Ch^-}, \quad f = \frac{f^*}{h^-}, \quad b = \frac{D}{C^2 \rho^-},$$
$$\nu = F^{-1} = \frac{(1-\rho) g h^-}{C^2}.$$

Then we obtain from (7)–(11)

$$(1 + f_x^2) f_{\psi\psi} + f_\psi^2 f_{xx} - 2 f_\psi f_x f_{x\psi} = 0 \tag{12}$$

for $0 < \psi < h$ and $-1 < \psi < 0$ with f defined as f^+ and f^-, respectively, in these two regions. At $\psi = 0$,

$$f^+ = f^- , \tag{13}$$

$$\nu f - \tau f_{xx}(1 + f_x^2)^{-3/2} + \frac{1}{2}\frac{1 + (f_x^-)^2}{(f_\psi^-)^2} - \frac{\rho}{2}\frac{1 + (f_x^+)^2}{(f_\psi^+)^2} = b; \tag{14}$$

at $\psi = -1$,

$$f^- = -1; \tag{15}$$

at $\psi = h$,

$$f^+ = h. \tag{16}$$

Now let $f = \psi + w(x, \psi)$. The equations (12)–(16) become

$$\begin{aligned} w_{\psi\psi} + w_{xx} = & - 2 w_\psi w_{xx} + 2 w_x w_{x\psi} - w_x^2 w_{\psi\psi} + 2 w_x w_\psi w_{x\psi} \\ & - w_\psi^2 w_{xx} = F_1(w), \end{aligned} \tag{17}$$

in $-1 < \psi < 0$ and $0 < \psi < h$, and w^+, w^- are defined in the same way as f^+ and f^-. At $\psi = 0$,

$$w^+ = w^-, \tag{18}$$

$$w_\psi^- - \rho w_\psi^+ - \nu w + \tau w_{xx} = \frac{3}{2}(w_\psi^-)^2 - \frac{3\rho}{2}(w_\psi^+)^2 + F_1^*(w), \tag{19}$$

where $b = (1 - \rho)/2$ and

$$\begin{aligned} F_1^*(w) = & (1 + w_x^2)^{3/2} \left(\frac{w_x^2}{2(1 + w_\psi^-)^2} - \frac{\rho w_x^2}{2(1 + w_\psi^+)^2} \right) \\ & + \left(\frac{\nu w - (2 w_\psi^- + (w_\psi^-)^2)}{2(1 + w_\psi^-)^2} + \frac{\rho(2 w_\psi^+ + (w_\psi^+)^2)}{2(1 + w_\psi^+)^2} \right) \\ & \times ((1 + w_x^2)^{3/2} - 1) + \frac{1}{2}(1 + w_\psi^-)^{-2} - \frac{1 - 2 w_\psi^- + 3(w_\psi^-)^2}{2} \\ & - \rho/2(1 + w_\psi^+)^{-2} + \frac{\rho(1 - 2 w_\psi^+ + 3(w_\psi^+)^2)}{2}; \end{aligned}$$

at $\psi = -1$,

$$w^- = 0; \tag{20}$$

at $\psi = h$,

$$w^+ = 0. \tag{21}$$

3. Formal derivation. Assume that w and ν have asymptotic expansions of the following form:

$$w = \epsilon w_1 + \epsilon^2 w_2 + \cdots , \tag{22}$$

$$\nu = \lambda_0 + \epsilon \lambda_1 . \tag{23}$$

Substitution of (22) and (23) with x replaced by $\epsilon^{-1/2}x$ in (17) to (21) yields the equations for the first approximation

$$\begin{aligned} &w_{1\psi\psi} = 0 \quad \text{in } -1 < \psi < 0, \qquad 0 < \psi < h, \\ &w_1^+ = w_1^-, \qquad w_{1\psi}^- - \rho w_{1\psi}^+ - \lambda_0 w_1^- = 0 \quad \text{at } \psi = 0, \\ &w_1^- = 0 \quad \text{at } \psi = -1, \quad w_1^+ = 0 \quad \text{at } \psi = h. \end{aligned}$$

By using these equations it is easy to get

$$w_1^+ = \eta(x)(\psi - h), \qquad w_1^- = -h\eta(x)(\psi + 1), \quad \lambda_0 = 1 + \frac{\rho}{h}.$$

For the second approximation from the equations,

$$\begin{aligned} &w_{2\psi\psi} = - w_{1xx} \quad \text{in } -1 < \psi < 0, \quad 0 < \psi < h, \\ &w_2^+ = w_2^-, w_{2\psi}^- - \rho w_{2\psi}^+ - \lambda_0 w_2^- = \lambda_1 w_1^- - \tau w_{1xx}^- \\ &\qquad + \frac{3}{2}(w_{1\psi}^-)^2 - \frac{3\rho}{2}(w_{1\psi}^+)^2 \quad \text{at } \psi = 0, \\ &w_2^- = 0 \quad \text{at } \psi = -1; \qquad w_2^+ = 0 \quad \text{at } \psi = h. \end{aligned}$$

By using the solvability condition for w_2, we can have the following equation for $\eta(x)$:

$$\left(\tau - \frac{1 + \rho h}{3}\right) \eta_{xx} - \lambda_1 \eta + \frac{3}{2}\left(h - \frac{\rho}{h}\right) \eta^2 = 0. \tag{24}$$

The solution of (24), which decays at infinity, is

$$\eta(x) = \left(\frac{\lambda_1 h}{h^2 - \rho}\right) \operatorname{sech}^2 \left(\left(\frac{\lambda_1}{\tau - \tau_0}\right)^{1/2} \frac{x}{2}\right), \tag{25}$$

if $\rho \neq h^2$ but not near h^2 and $\tau \neq \tau_0 = (1 + \rho h)/3$ but not near τ_0. In the following we shall show that (25) indeed yields an asymptotic approximation to the solution of (17)–(21) when ν is sufficiently close to $1 + (\rho/h)$, $\rho \neq h^2$ but not near h^2, and $\tau > \tau_0$ but not near τ_0.

4. Existence theorem.

4.1. Preliminaries. In the following we always assume that $\rho \neq h^2$ but not near h^2, and $\tau > \tau_0$ but not near τ_0 unless specified otherwise. Let

$$\begin{aligned} &\nu = \lambda_0 + \lambda_1 \epsilon = \left(1 + \frac{\rho}{h}\right) + \lambda_1 \epsilon, \qquad \lambda_1 > 0, \\ &w^+(x, \psi) = \epsilon(\eta(x)(\psi - h) + \theta^+(x, \psi)) \quad \text{for } 0 < \psi < h, \\ &w^-(x, \psi) = \epsilon(\eta(x)(\psi + 1)(-h) + \theta^-(x, \psi)) \quad \text{for } -1 < \psi < 0. \end{aligned}$$

Substitute these expressions for w^+ and w^- in (17)–(21) and obtain

$$\theta^+_{\psi\psi} + \theta^+_{xx} = -\eta_{xx}(\psi - h) + \frac{1}{\epsilon}F_1(w^+) = F_2^+(x,\psi), \qquad 0 < \psi < h, \tag{26}$$

$$\theta^-_{\psi\psi} + \theta^-_{xx} = h\eta_{xx}(\psi + 1) + \frac{1}{\epsilon}F_1(w^-) = F_2^-(x,\psi), \qquad -1 < \psi < 0; \tag{27}$$

at $\psi = 0$,

$$\theta^+(x,0) = \theta^-(x,0), \tag{28}$$

$$\begin{aligned} \theta^-_\psi - \rho\theta^+_\psi - \lambda_0\theta^- + \tau\theta_{xx} &= \tau h\eta_{xx} + \lambda_1\epsilon(-h\eta + \theta^-) \\ &\quad + \frac{3}{2}\epsilon(-h\eta + \theta^-_\psi)^2 - \frac{3\rho}{2}\epsilon(\eta + \theta^+_\psi)^2 + F_1^*(w)\epsilon^{-1} \\ &= \tau h\eta_{xx} - \lambda_1 h\eta\epsilon + \frac{3}{2}\epsilon h^2\eta^2 - \frac{3\rho}{2}\epsilon\eta^2 + F_2^*(\eta,\theta) \\ &= G_1(x); \end{aligned} \tag{29}$$

at $\psi = -1$,

$$\theta^- = 0; \tag{30}$$

at $\psi = h$,

$$\theta^+ = 0\,. \tag{31}$$

Since

$$j(\psi) = \begin{cases} \psi - h, & 0 < \psi < h, \\ (\psi + 1)(-h), & -1 < \psi < 0, \end{cases}$$

is the nontrivial solution of the homogeneous equations (26)–(31); to solve θ from (26)–(31) we impose the following solvability condition:

$$\int_{-1}^{0} F_2^-(x,\psi)(-h)(\psi + 1)d\psi + \rho\int_0^h F_2^+(x,\psi)(\psi - h)d\psi + hG_1(x) = 0, \tag{32}$$

which was used in the formal derivation of (24). As we shall see later, this condition excludes the function $j(\psi)$ in the eigenfunction expansion of $\theta(x,\psi)$. From (17), (26), (27), (29), and (32), we obtain an equation for η,

$$\begin{aligned} &\left(\tau - \frac{1+\rho h}{3}\right)\eta_{xx} - \lambda_1\eta\epsilon + \frac{3h}{2}\epsilon\eta^2 - \frac{3\rho}{2h}\epsilon\eta^2 \\ &= \frac{1}{h^2}\left(-F_2^*(\eta,\theta) - \int_{-1}^{0}\frac{1}{\epsilon}F_1(w^-)(-h)(\psi + 1)d\psi - \rho\int_0^h \frac{1}{\epsilon}F_1(w^+)(\psi - h)d\psi\right) \\ &= \Psi^*(\eta,\theta). \end{aligned} \tag{33}$$

Since

$$S(x) = \left(\frac{\lambda_1 h}{h^2 - \rho}\right)\operatorname{sech}^2\left(\left(\frac{\lambda_1}{\tau - \tau_0}\right)^{1/2}\frac{\epsilon^{1/2}x}{2}\right) \tag{34}$$

is a solution of (33) with $\Psi^*(\eta,\theta)=0$, we write

$$\eta(x)=S(x)+\omega(x).$$

Then from (33), we have

$$\begin{aligned}(\tau-\tau_0)\omega_{xx}-\lambda_1\omega\epsilon+3\left(h-\frac{\rho}{h}\right)\epsilon S\omega \\ =-\frac{3}{2}\left(h-\frac{\rho}{h}\right)\epsilon\omega^2+\Psi^*(S+\omega,\theta) \\ =\Psi(\omega,\theta)(\tau-\tau_0).\end{aligned} \tag{35}$$

Also from (26), (27), and (29), we write

$$\begin{aligned}&F_2^+(x,\psi)=F^+(\omega,\theta),F_2^-(x,\psi)=F^-(\omega,\theta),\\ &G_1(x)=G(\omega,\theta).\end{aligned}$$

Our goal is to find a small solution of (26)–(31) and (35) when ϵ is small. Before we prove the existence theorem, let us define some Banach spaces to be used later.

We denote $H^s(\mathbf{R})$ as the usual Sobolev space on $\mathbf{R}$ with $s\geq 0$,

$$\begin{aligned}X^s=&\{f(x)\in L^2(\mathbf{R},\mathbf{R})\mid f(x)\cosh(\mu\epsilon^{1/2}x)\in H^s(\mathbf{R})\ \text{ with } f \text{ even and}\\ &\|f\|_{X^s}=\|f(x)\cosh(\mu\epsilon^{1/2}x)\|_{H^s(\mathbf{R})}\},\end{aligned}$$

where μ is a small constant to be determined later, and

$$\begin{aligned}Y^s=\Bigg\{&f(x,y)\in L^2(\mathbf{R}\times((-1,0)\cup(0,h)))\mid f(\cdot,y)\in X^s \text{ for almost all}\\ &y \text{ and } \|f\|_{Y^s}^2=\int_{-1}^{h}\|f\|_{X^s}^2dy<+\infty\Bigg\}.\end{aligned}$$

Define for $s\geq 2$,

$$\begin{aligned}B^s=\Bigg\{&f(x,y)\in L^2(\mathbf{R}\times((-1,0)\cup(0,h)))\ \Bigg|\ \frac{\partial^m f(x,y)}{\partial y^m}\in Y^{s-m},f^+(x,0)=f^-(x,0),\\ &\text{for } 0\leq m\leq 2,\left.\frac{\partial f^\pm(x,y)}{\partial y}\right|_{y=0}\in X^{s-2},\quad \text{with norm}\\ &\|f\|_{B^s}=\sum_{j=0}^{2}\{\|D^jf\|_{Y^{s-2}}+\|D_x^jf(\cdot,0)\|_{X^{s-2}}\}+\|f_y^\pm(\cdot,0)\|_{X^{s-2}}<+\infty\Bigg\},\end{aligned}$$

where $B^s\supset X^s$. Now we study (26)–(31) and (35) separately.

4.2. Auxiliary condition. First we take up (35). Let

$$\begin{aligned}\mathcal{L}(\omega)&=\omega_{xx}-\frac{\lambda_1}{\beta}\omega\epsilon+3\left(h-\frac{\rho}{h}\right)\epsilon\frac{S\omega}{\beta}\\ &=\Psi\quad \text{for } 0<x<+\infty,\end{aligned} \tag{36}$$

$$\omega_x = 0 \quad \text{at} \ x = 0, \quad \text{where } \beta = \tau - \tau_0.$$

The homogeneous equation $\mathcal{L}(\omega) = 0$ possesses two linearly independent solutions

$$\varphi(x) = -\left(\frac{\lambda_1^3}{\beta}\right)^{1/2} \operatorname{sech}^2 t \tanh t,$$

$$\psi(x) = -2\beta(\lambda_1^2 \epsilon^{1/2})^{-1}(\operatorname{sech}^2 t)(\tanh t)\left(5\left(\frac{3t}{8} + \frac{3}{8}(\sinh t)\cosh t + \frac{1}{4}(\sinh t)\cosh^3 t\right) - (\cosh^5 t)\sinh^{-1} t\right),$$

where $t = \frac{1}{2}(\lambda_1\epsilon/\beta)^{1/2}x$, $\varphi(x)$ and $\psi(x)$ are, respectively, odd and even functions. Assume $f(x)$ is bounded and even. Then

$$\begin{aligned} &\mathcal{L}(\omega) = f, \qquad 0 < x < +\infty, \\ &\omega_x = 0 \quad \text{at} \ x = 0, \end{aligned} \tag{37}$$

has the solution

$$\begin{aligned} \omega(x) &= \varphi(x)\int_0^x \psi(s)f(s)ds + \psi(x)\int_x^{+\infty} \varphi(s)f(s)ds \ \text{ for } x \geq 0 \\ &= \int_0^{+\infty} k^*(x,s)f(s)ds, \end{aligned} \tag{38}$$

and

$$\omega(x) = \int_0^{+\infty} k^*(-x,s)f(s)ds \quad \text{for } x < 0.$$

In the following we prove two lemmas for the estimates of $\omega(x)$.

LEMMA 1. *If $f(x) \in X^s$, $s \geq 0$ and μ is small, then*

$$\|\omega(x)\|_{X^{s+2}} \leq K\epsilon^{-1}\|f(x)\|_{X^s},$$

where K is independent of ϵ and $f(x)$.

Proof. Here we only prove the case $s = 0$. The proofs of other cases are the same. From (38) we know that if $f(x)$ is even; $\omega(x)$ is also even; therefore,

$$\begin{aligned} \|\omega(x)\cosh(\mu\epsilon^{1/2}x)\|^2_{L^2(\mathbf{R})} &\leq K\int_0^{+\infty} |\omega(x)\cosh(\mu\epsilon^{1/2}x)|^2 dx \\ &\leq K\int_0^{+\infty}\left(\int_0^{+\infty} |k^*(x,s)\cos(\mu\epsilon^{1/2}x)f(s)|ds\right)^2 dx \\ &= K\int_0^{+\infty}\left(\int_0^{+\infty} k'(x,s)g(s)ds\right)^2 dx = I, \end{aligned}$$

where

$$\begin{aligned} k'(x,s) &= |k^*(x,s)|\cosh(\mu\epsilon^{1/2}x)\cosh^{-1}(\mu\epsilon^{1/2}s), \\ g(s) &= f(s)\cosh(\mu\epsilon^{1/2}s), \end{aligned}$$

and

$$\begin{aligned} I &\le K\int_0^{+\infty}\int_0^{+\infty} k'(x,t)dt\int_0^{+\infty} k'(x,s)g^2(s)dsdx \\ &\le K\left(\max\left(\sup_s\int_0^{\infty} k'(x,s)dx, \sup_x\int_0^{+\infty} k'(x,s)ds\right)\right)^2 \|g(s)\|_{L^2}^2 . \end{aligned}$$

But

$$\begin{aligned} \int_0^{+\infty} k'(x,s)dx &\le K\Bigg(\int_0^{s} |\varphi(s)\psi(x)\mid \exp(\mu\epsilon^{1/2}(x-s))dx \\ &\qquad + \int_s^{+\infty} |\varphi(x)\psi(s)| \exp(\mu\epsilon^{1/2}(x-s))dx\Bigg) \\ &\le K\epsilon^{-1/2}\Bigg(\int_0^{s} \exp\left(\frac{\lambda_1\epsilon}{\beta}^{1/2}\frac{1}{2}(x-s) - \mu\epsilon^{1/2}(x-s)\right) dx \\ &\qquad + \int_s^{+\infty} \exp\left(\frac{\lambda_1\epsilon}{\beta}^{1/2}\frac{1}{2}(s-x) - \mu\epsilon^{1/2}(x-s)\right) dx\Bigg) \\ &\le K\epsilon^{-1} , \end{aligned}$$

if $0 < \mu < \frac{1}{2}(\lambda_1/\beta)^{1/2}$. The same is true for $\int_0^{+\infty} k'(x,s)ds$. Thus

$$\|\omega(x)\|_{X^0} \le K\epsilon^{-1}\|f(x)\|_{X^0} .$$

Also

$$\|\omega\|_{X^1}^2 \le K(\|\omega'(x)\cosh(\mu\epsilon^{1/2}x)\|_{L^2(\mathbf{R})}^2 + \|\omega\|_{X^0}^2) ,$$

and we only need to get the estimate for $\|\omega'(x)\cosh(\mu\epsilon^{1/2}x)\|_{L^2(\mathbf{R})}^2$. But from (38) we have

$$\begin{aligned} \omega'(x) &= \varphi'(x)\int_0^{x} \psi(s)f(s)ds + \psi'(x)\int_x^{+\infty} \varphi(s)f(s)ds \quad \text{for } x \ge 0 \\ &= \int_0^{+\infty} k_1^*(x,s)f(s)ds, \end{aligned}$$

where $\omega'(x)$ is odd and $k_1^*(x,s)$ has the same asymptotic behavior as $k^*(x,s)$, where x or s is large. Therefore, by the same argument, we have

$$\|\omega(x)\|_{X^1} \le K\epsilon^{-1}\|f\|_{X^0} .$$

However, $\omega(x)$ satisfies (37), and from (37) and the estimates of $\omega(x)$ in X^1, we have

$$\|\omega(x)\|_{X^2} \le K\epsilon^{-1}\|f\|_{X^0} ,$$

if $0 < \mu < (\lambda_1/4\beta)^{1/2}$. Also we have the following lemma, the proof of which is omitted.

LEMMA 2. *If* $u_1 \in H^s(\mathbf{R})$, $u_2 \in X^s$, $s \ge 1$, *then*

$$\|u_1u_2\|_{X^s} \le K\|u_1\|_{H^s}\|u_2\|_{X^s}.$$

Then by checking term by term in $\Psi(w,\theta)$ carefully and using Lemmas 1 and 2, we have the following.

THEOREM 1. *If $\omega, \theta \in B^s$ and $\|\omega\|_{B^s} = \|\omega\|_{X^s} \leq K\epsilon^{1/2}$, $\|\theta\|_{B^s} \leq K\epsilon$ (or $K\epsilon^{1/2}$) with $s \geq 3$, then $\Psi(\omega, \theta) \in X^{s-2}$,*

$$\|\Psi(\omega,\theta)\|_{X^{s-2}} \leq K\epsilon^2 \quad (or\ K\epsilon^{3/2}),$$

and

$$\|\mathcal{L}^{-1}\Psi(\omega,\theta)\|_{X^s} \leq K\epsilon \quad (or\ K\epsilon^{1/2}),$$

where

$$\mathcal{L}^{-1}\Psi(\omega,\theta) = \begin{cases} \displaystyle\int_0^{+\infty} k^*(x,s)\Psi(\omega,\theta)ds, & x \geq 0, \\ \displaystyle\int_0^{+\infty} k^*(-x,s)\Psi(\omega,\theta)ds, & x < 0. \end{cases}$$

4.3. Partial differential equations. Now we consider (26)–(31). First let us discuss the following equations:

(39) $$u^-_{xx} + u^-_{\psi\psi} = \varphi^-_1(x,\psi), \qquad -\infty < x < +\infty, \quad -1 < \psi < 0,$$

(40) $$u^+_{xx} + u^+_{\psi\psi} = \varphi^+_1(x,\psi), \qquad -\infty < x < +\infty, \quad 0 < \psi < h;$$

at $\psi = 0$,

(41) $$u^+(x,0) = u^-(x,0),$$

(42) $$\tau^{-1}(u^-_\psi - \rho u^+_\psi - \lambda_0 u^-) + u^-_{xx} = \varphi_2(x);$$

at $\psi = -1$,

(43) $$u^- = 0;$$

at $\psi = h$,

(44) $$u^+ = 0.$$

Here $\varphi_1(x,\psi) \in B^s$ and $\varphi_2(x) \in X^s$ for $s \geq 2$. Following Beale [8], we make use of the family of eigenfunctions of ψ satisfying

(45) $$v_{\psi\psi} - \sigma v = 0 \quad \text{in } -1 < \psi < 0 \quad \text{and} \quad 0 < \psi < h\,,$$

(46) $$v^+ = v^-, v^-_\psi - \rho v^+_\psi - \lambda_0 v^- - \tau\sigma v^- = 0 \quad \text{at } \psi = 0,$$

(47) $$v^- = 0 \quad \text{at } \psi = -1; \qquad v^+ = 0 \quad \text{at } \psi = h\,.$$

To discuss the completeness of the eigenfunctions of (45)–(47), first let us set up some Hilbert space. Let $(f(\psi), q) \in (L^2(-1,h)) \times \mathbf{R}$, where q is a constant and define

(48) $$\langle (f,q),(g,r)\rangle = \int_{-1}^0 f^-g^-d\psi + \rho\int_0^h f^+g^+d\psi - \tau qr,$$

(49) $$|(f,q)|^2 = \langle (f,q),(f,q)\rangle\,.$$

Then define

$$H = \left\{ (f,q) \in (L^2(-1,h)) \times \mathbf{R} \,\middle|\, \int_{-1}^{0} f^-(\psi)(-h)(\psi+1)d\psi \right.$$
$$+ \rho \int_0^h f^+(\psi)(\psi - h)d\psi + \tau h q = 0 \text{ and the norm is}$$
$$\left. \text{defined by (49)} \right\}.$$

We show that $\langle (f,q),(g,r)\rangle$ is an inner product. Since the proofs of other conditions are straightforward, the only thing that we need to prove is

$$\langle (f,q),(f,q)\rangle > 0,$$

if $0 \neq (f,q) \in H$. Let $u^* = -h(\psi+1)$ for $-1 < \psi < 0$, $u^* = \rho(\psi - h)$ for $0 < \psi < h$. Then if $(f,q) \in H$,

$$q = \frac{\int_{-1}^h f u^* d\psi}{-h\tau}.$$

So

$$q^2 \le (h\tau)^{-2} \left(\int_{-1}^0 (f^-)^2 d\psi + \int_0^h (\rho^{1/2} f^+)^2 d\psi \right)$$
$$\times \left(\int_{-1}^0 h^2(\psi+1)^2 + \rho \int_0^h (\psi - h)^2 d\psi \right)$$
$$= \tau^{-2}(1+\rho h) \frac{\int_{-1}^0 (f^-)^2 d\psi + \rho \int_0^h (f^+)^2 d\psi}{3}.$$

Therefore, by (48) and (49),

$$|(f,q)|^2 \ge (1 - (3\tau)^{-1}(1+\rho h)) \left(\int_{-1}^0 (f^-)^2 d\psi + \rho \int_0^h (f^+)^2 d\psi \right).$$

Thus if $\tau > \tau_0$, $|(f,q)| > 0$ unless $(f,q) = 0$. Hence H is a Hilbert space in reference to the inner product (48). Note that $|(f,q)|$ is equivalent to the norm $\|(f,q)\| = (\|f\|_{L^2}^2 + |q|^2)^{1/2}$ for $(f,q) \in H$.

LEMMA 3. *The space H is a Hilbert space.*

The condition imposed on any two functions in H is the orthogonal condition with the eigenfunction of (45)–(47) corresponding to $\sigma = 0$ under the inner product $\langle \cdot , \cdot \rangle$, and H is a closed subspace in $(L^2(-1,h) \times \mathbf{R})$. Hence H is a Hilbert space.

Next, we define an operator in H:

$$\mathcal{A}(u,q) = \left(u_{\psi\psi}, \tau^{-1}(u_\psi^- - \rho u_\psi^+ - \lambda_0 u^-)\Big|_{\psi=0} \right), \tag{50}$$

with

$$D(\mathcal{A}) = \{(u,q) \in H \mid u^- \in H^2(-1,0), u^+ \in H^2(0,h),$$
$$u^-(0) = u^+(0) = q, u^-(-1) = 0 = u^+(h)\}.$$

Obviously the solutions of (45)–(47) correspond to the eigenfunctions of $\mathcal{A}$ in H with eigenvalue σ and $\mathcal{A}$ closed and densely defined. The equality
(51)

$$\begin{aligned}\langle \mathcal{A}(u,q),(v,r)\rangle &= \int_{-1}^{0} u^{-}_{\psi\psi}v^{-}d\psi + \rho\int_{0}^{h} u^{+}_{\psi\psi}v^{+}d\psi - r(u^{-}_{\psi} - u^{+}_{\psi} - \lambda_0 u^{-})v^{+}\Big|_{\psi=0} \\ &= -\left(\int_{-1}^{0} u^{-}_{\psi}v^{-}_{\psi}d\psi + \rho\int_{0}^{h} u^{+}_{\psi}v^{+}_{\psi}d\psi\right) + \lambda_0 u^{-}(0)v^{-}(0),\end{aligned}$$

for $(u,q),(v,r) \in D(\mathcal{A})$, shows that $\mathcal{A}$ is symmetric with respect to $\langle\ \cdot\ ,\ \cdot\ \rangle$, and implies that all eigenvalues of $\mathcal{A}$, if they exist, must be real. By (51) and the Schwarz inequality,

$$\langle \mathcal{A}(u,q),(u,q)\rangle = -\left(\int_{-1}^{0}(u^{-}_{\psi})^2 d\psi + \rho\int_{0}^{h}(u^{+}_{\psi})^2 d\psi\right) + \lambda_0(u^{-}(0))^2 < 0,$$

if $(u,q) \in D(\mathcal{A})$ since

$$\begin{aligned}(u^{-}(0))^2 &= \left(\int_{-1}^{0} u^{-}_{\psi}d\psi\right)^2 \leq \int_{-1}^{0}(u^{-}_{\psi})^2 d\psi, \\ (u^{+}(0))^2 &= \left(\int_{0}^{h} u^{+}_{\psi}d\psi\right)^2 \leq h\int_{0}^{h}(u^{+}_{\psi})^2 d\psi,\end{aligned}$$

and $\lambda_0 = 1 + (\rho/h)$, where the equalities hold if and only if $u = j(\psi)$ defined in (32), which is not in H. This shows that all eigenvalues of $\mathcal{A}$ are nonpositive. From (51), H is invariant under $\mathcal{A}$. In order to show that the eigenfunctions of $\mathcal{A}$ is a basis of H, it suffices to show that $(\mathcal{A} - \sigma)^{-1}$ exists and is compact in H for some $\sigma > 0$. But $(\mathcal{A} - \sigma)(u, u(0)) = (f, q)$ is equivalent to solve

$$\begin{aligned}&u_{\psi\psi} - \sigma u = f \quad \text{in} \quad -1 < \psi < 0 \quad \text{and} \quad 0 < \psi < h, \\ &u^{+}(0) = u^{-}(0), u^{-}_{\psi} - \rho u^{+}_{\psi} - \lambda_0 u^{-} - \tau\sigma u^{-} = \tau q \quad \text{at} \ \psi = 0, \\ &u^{-} = 0 \quad \text{at} \ \psi = -1; \qquad u^{+} = 0 \quad \text{at} \ \psi = h.\end{aligned}$$

Then the solution of the above equations is easily obtained as follows:

$$\begin{aligned}u^{+} = -\frac{a(\psi - h)}{h} + \Bigg(&\sinh\sigma^{1/2}(\psi - h)\int_{0}^{\psi}\sinh\sigma^{1/2}\psi\left(f^{+} - \frac{a\sigma(\psi - h)}{h}\right)d\psi \\ &+ \sinh\sigma^{1/2}\psi\int_{\psi}^{h}\sinh\sigma^{1/2}(\psi - h)\left(f^{+} - \frac{a\sigma(\psi - h)}{h}\right)d\psi\Bigg) \\ &\Big/(\sigma^{1/2}\sinh\sigma^{1/2}h),\end{aligned}$$

$$\begin{aligned}u^{-} = a(\psi + 1) + \Bigg(&\sinh\sigma^{1/2}\psi\int_{-1}^{\psi}\sinh\sigma^{1/2}(\psi + 1)(f^{-} + a\sigma(\psi + 1))d\psi \\ &+ \sinh\sigma^{1/2}(\psi + 1)\int_{\psi}^{0}\sinh\sigma^{1/2}\psi(f^{-} + a\sigma(\psi + 1))d\psi\Bigg) \\ &\Big/(\sigma^{1/2}\sinh\sigma^{1/2}),\end{aligned}$$

where

$$a = \Gamma^{-1}(\sigma)\left(\tau q - (\sinh\sigma^{1/2})^{-1}\int_{-1}^{0}\sinh\sigma^{1/2}(\psi+1)f^{-}d\psi\right.$$
$$\left.+\rho(\sinh\sigma^{1/2}h)^{-1}\int_{0}^{h}\sinh\sigma^{1/2}(\psi-h)f^{+}d\psi\right),$$

and

$$\Gamma(\sigma) = \sigma^{1/2}(\mathrm{tanh}\sigma^{1/2})^{-1} + \rho\sigma^{1/2}(\mathrm{tanh}\sigma^{1/2}h)^{-1} - \lambda_0 - \tau\sigma \neq 0$$

for all $\sigma > 0$ as shown in the Appendix. Obviously $(u, u(0))$ constructed in this way is in H if $(f, q) \in H$. By standard arguments it is not difficult to show that $(\mathcal{A}-\sigma)^{-1}$ on H is compact. Furthermore, it follows that there are countably infinitely many eigenvalues, and the range of $(\mathcal{A}-\sigma)^{-1}$ is $D(\mathcal{A})$. Also in the Appendix the asymptotic behavior of these eigenvalues is given.

We summarize the results as follows.

LEMMA 4. *The system* (45)–(47) *has solutions* $v = v_n$, $\sigma = \sigma_n$, $n = 1, 2, \ldots$, *where* $(v_n(\psi), v_n(0)) \in H$, $\sigma_n < 0$. *The vectors* $(v_n(\psi), v_n(0))$, $n = 1, 2, \ldots$, *form a basis of* H *and are orthonormal with respect to the inner product* (50). *Here* $|\sigma_n| \cong K_1 n^2$ *for large* n *where* $K_1 \geq c > 0$ *with* c *fixed.*

Now we can deal with (39)–(44). Assume that $(\varphi_1(x,\psi), \varphi_2(x)) \in H$ for all $x \in \mathbf{R}$. By Lemma 4, we have

$$(\varphi_1(x,\psi), \varphi_2(x)) = \sum_{n=1}^{\infty} a_n(x)(v_n(\psi), v_n(0)),$$

where the convergence is under the norm (49) and

$$(52)\qquad a_n(x) = \int_{-1}^{0}\varphi_1^{-}(x,\psi)v_n^{-}(x)d\psi + \rho\int_{0}^{h}\varphi_1^{+}(x,\psi)v_n^{+}(x)d\psi - \tau\varphi_2(x)v_n(0).$$

Multiplying (39) and (40) on both sides by $v_n(\psi)$ and integrating by parts twice we have

$$(53)\qquad C_{nxx}(x) + \sigma_n C_n(x) = a_n(x),$$

where

$$C_n = \int_{-1}^{0}u^{-}(x,\psi)v_n^{-}(\psi)d\psi + \rho\int_{0}^{h}u^{+}(x,\psi)v_n^{+}(\psi)d\psi - \tau u^{-}(x,0)v_n^{-}(0).$$

If $a_n(x)$ and $C_n(x)$ are even, by finding the Green's function of (53) and following the same proof as in Lemma 1, we have the following.

LEMMA 5. *If* $a_n(x) \in X^s$ *and* $C_n(x)$ *satisfies* (53), *then* $C_n(x) \in X^{s+2}$,

$$\|C_n\|_{X^{s+i}} \leq K|\sigma_n|^{-(2-i)/2}\|a_n(x)\|_{X^s} \quad \textit{for } i = 0, 1, 2.$$

Therefore,

$$(54)\qquad (u(x,\psi), u(x,0)) = \sum_{n=1}^{\infty} C_n(x)(v_n(\psi), v_n(0))$$

is in the Hilbert space H, and

$$(55)\qquad \sum_{n=1}^{\infty} \|C_n(x)\|^2_{X^{s+i}} \le K \sum_{n=1}^{\infty} |\sigma_n|^{-(2-i)} \|a_n(x)\|^2_{X^s} \quad \text{for } i = 0, 1, 2.$$

If $(u(x,\psi), u(x,0))$ is defined by (54), by (55) we have

$$\sum_{i=0}^{2} (\|D^i_x u(x,\psi)\|_{Y^s} + \|D^j_x u(x,0)\|_{X^s}) \le K(\|\varphi_1(x,\psi)\|_{Y^s} + \|\varphi_2\|_{X^s}).$$

Since $(u(x,\psi), u(x,0))$ satisfies (39)–(44); so by (39) and (40)

$$\|u_{\psi\psi}\|_{Y^s} \le K(\|\varphi_1\|_{Y^s} + \|\varphi_2\|_{X^s}),$$

and by usual interpolation theorems, the same is true of u_ψ. Then by using $u = 0$ on $\psi = -1$ and h and $u(x,0) \in X^s$, we can easily show that

$$u_\psi(x,\psi)\Big|_{\psi=\psi_s} \in X_s$$

and

$$\|u_\psi(x,\psi_s)\|_{X^s} \le K(\|\varphi_1\|_{Y^s} + \|\varphi_2\|_{X^s}),$$

where ψ_s is fixed and when $\psi_s \to 0$, $u_\psi(x,\psi_s)$ in the limit corresponds to $u^+_\psi(x,0)$, and $u^-_\psi(x,0)$. So

$$\|u^\pm_\psi(x,0)\|_{X^s} \le K(\|\varphi_1\|_{Y^s} + \|\varphi_2\|_{X^s}).$$

If $\varphi_1(x,\psi)$ and $\varphi_2(x)$ have continuous derivatives at least up to $2s+2$ order with respect to x and ψ and $(\varphi_1, \varphi_2) \in H$, then we may multiply (39) and (40) by u_{xx} and integrate by parts several times to obtain

$$(56)\qquad \|u_{x\psi}\|_{Y^s} \le K(\|\varphi_1\|_{Y^s} + \|\varphi_2\|_{X^s}).$$

For the general $\varphi_1(x,\psi) \in Y^s$ and $\varphi_2 \in X^s$ with $(\varphi_1, \varphi_2) \in H$ we can construct a sequence of smooth functions in H approaching (φ_1, φ_2), and by usual techniques in elliptic operator theory we get the a priori estimate (56).

Now combining all estimates obtained we have

$$(57)\qquad \|u\|_{B^{s+2}} \le K(\|\varphi_1\|_{Y^s} + \|\varphi_2\|_{X^s}),$$

if $\varphi_1 \in Y^s$ and $\varphi_2 \in X^s$ with $(\varphi_1, \varphi_2) \in H$. Therefore we summarize the above results as follows.

THEOREM 2. *If $\varphi_1 \in Y^s$, $\varphi_2 \in X^s$ and $(\varphi_1, \varphi_2) \in H$, then the solution $u(x,\psi)$ of* (39)–(44) *exists in B^{s+2}, is unique in H, and satisfies* (57).

Now we use the following notation to express the solution of (39)–(44) in terms of $(\varphi_1(x,\psi), \varphi_2(x))$,

$$(58)\qquad u(x,\psi) = \mathcal{P}^{-1}(\varphi_1(x,\psi), \varphi_2(x)).$$

Let us go back to (26)–(31). Write

$$\begin{aligned} F_2(x,\psi) &= F^*(\omega,\theta) = M_1(\omega) + \tilde{F}_2(\omega,\theta), \\ G_1(x) &= G^*(\omega,\theta) = M_2(\omega) + \tilde{G}_1(\omega,\theta), \end{aligned}$$

where $M_1(\omega)$ and $M_2(\omega)$ contain only the terms with $\omega_{xx}, \epsilon\omega$, and $\epsilon S\omega$, which appear in F_2 and G_1, respectively, and $\tilde{F}_2$ and $\tilde{G}_1$ are the remainders. By the derivation of (35) from (32), we can see that the left-hand side of (35) denoted as $(\tau - \tau_0)\mathcal{L}(\omega)$ in (36) is from (M_1, M_2) only, and the right-hand side of (35) is from $(\tilde{F}_2, \tilde{G}_1)$ and the left-hand side of (32) is equal to $(\tau - \tau_0)(\mathcal{L}(\omega) - \Psi)$. Since $\mathcal{L}^{-1}$ exists, define

$$\begin{aligned} F(\omega,\theta) &= M_1(\mathcal{L}^{-1}(\Psi(\omega,\theta))) + \tilde{F}_2(\omega,\theta), \\ G(\omega,\theta) &= M_2(\mathcal{L}^{-1}(\Psi(\omega,\theta))) + \tilde{G}_1(\omega,\theta), \end{aligned}$$

which is equivalent to (F_2, G_1) for this problem. If we use this (F, G) in (32), then $(F(\omega,\theta), G(\omega,\theta)) \in H$ for all ω, and θ since by a similar derivation of (35) from (32) and noting that $\mathcal{L}^{-1}$ only appears in (M_1, M_2) we have

$$\begin{aligned} &\int_{-1}^{0} F^-(\omega,\theta)(-h)(\psi+1)d\psi + \rho\int_0^h F^+(\omega,\theta)(\psi - h)d\psi + \tau h G(\omega,\theta) \\ &\quad = (\tau - \tau_0)(\mathcal{L}(\mathcal{L}^{-1}(\Psi(\omega,\theta))) - \Psi(\omega,\theta)) = 0. \end{aligned}$$

If we assume that $\theta(x,\psi) \in B^{s+2}$ and $\omega \in X^{s+2}$, then by checking the terms in F^+, F^- and G very carefully and using Lemma 2 and Theorem 1, we have $F \in Y^s$, $G \in X^s$, and

$$\|F\|_{Y^s} \le K\epsilon(\|\theta\|_{B^{s+2}} + \|\omega\|_{X^{s+2}}), \tag{59}$$

$$\|G\|_{X^s} \le K\epsilon(\|\theta\|_{B^{s+2}} + \|\omega\|_{X^{s+2}}). \tag{60}$$

By Theorem 2, we obtain the following.

THEOREM 3. *If $\theta(x,\psi) \in B^{s+2}$ and $\omega(x) \in X^{s+2}$ are bounded with respect to their respective norms, then the solution of* (26)–(31) *with F_2 and G_1 replaced by F and G, respectively, exists and satisfies*

$$\|\mathcal{P}^{-1}(F,G)\|_{B^{s+2}} \le K\epsilon(\|\theta\|_{B^{s+2}} + \|\omega\|_{X^{s+2}}).$$

4.4. The existence proof. Now we need to solve (26)–(31) and (35), which is equivalent to (32). We know that

$$(F_2(x,\psi), G_1(x)) = (F^*(\omega,\theta), G^*(\omega,\theta)) \in H$$

since (32) implies that (F_2, G_1) satisfies the orthogonal condition defined for the space H. Convert (35) to

$$\omega = \mathcal{L}^{-1}(\Psi(\omega,\theta)), \tag{61}$$

and (26) to (31) to

$$\theta = \mathcal{P}^{-1}(F_2(\omega,\theta), G_1(\omega,\theta)). \tag{62}$$

Then instead of solving (61) and (62) together, we solve the following equivalent equations:

$$\theta = \mathcal{P}^{-1}(F(\omega,\theta), G(\omega,\theta)) = \mathcal{T}_1(\theta,\omega), \tag{63}$$

$$\omega = \mathcal{L}^{-1}(\Psi(\omega, \mathcal{P}^{-1}(F(\omega,\theta), G(\omega,\theta)))) = \mathcal{T}_2(\theta,\omega), \tag{64}$$

since we need to use this substitution once to meet the condition in Theorem 1 and $(F, G) \in H$ to satisfy the condition in Theorem 3. Now we need to prove that when ϵ is small, (63) and (64) possess a fixed point. First we define a closed convex set in the Banach space $B^{s+2} \times X^{s+2}$ for $s \geq 2$,

$$\mathcal{S}_b = \{Z = (\theta,\omega) \in B^{s+2} \times X^{s+2} \mid \quad |||Z||| = \|\theta\|_{B^{s+2}} + \|\omega\|_{X^{s+2}} \leq b\epsilon^{1/2}\}.$$

Let

$$\mathcal{T}(\theta,\omega) = (\mathcal{T}_1(\theta,\omega), \mathcal{T}_2(\theta,\omega)).$$

Then we try to find a fixed point of the operator $\mathcal{T}$ in $\mathcal{S}_b$. From Theorems 1 and 3 it follows that $\mathcal{T}$ maps $\mathcal{S}_b$ into itself if $K\epsilon^{1/2} < b$, which is always possible if ϵ is small. Also by using similar proofs as in Theorem 1 and 3 we have

THEOREM 4. *If* $Z^{(1)} = (\theta^{(1)}, \omega^{(1)})$ *and* $Z^{(2)} = (\theta^{(2)}, \omega^{(2)}) \in \mathcal{S}_b$, *then*

$$|||\mathcal{T}(Z^{(1)}) - \mathcal{T}(Z^{(2)})||| \leq K\epsilon^{1/2}|||Z^{(1)} - Z^{(2)}|||.$$

Now we may choose $K\epsilon^{1/2} \leq \frac{1}{2}$ in Theorem 4 for smaller ϵ so that $\mathcal{T}$ maps $\mathcal{S}_b$ into $\mathcal{S}_b$ and is a contraction in $\mathcal{S}_b$. By contraction mapping theorem, there exists a unique point (θ_0, ω_0) so that

$$\theta_0 = \mathcal{P}^{-1}(F(\omega_0,\theta_0), G(\omega_0,\theta_0)),$$
$$\omega_0 = \mathcal{L}^{-1}(\Psi(\omega_0,\theta_0)),$$

or equivalently,

$$\theta_0 = \mathcal{P}^{-1}(F^*(\omega_0,\theta_0), G^*(\omega_0,\theta_0)),$$
$$\omega_0 = \mathcal{L}^{-1}(\Psi(\omega_0,\theta_0)).$$

Finally we have the following theorem.

THEOREM 5. *If* $\nu = (1 + (\rho/h)) + \lambda_1\epsilon$, $\lambda_1 > 0$, *then for small* $\epsilon > 0$, *there exists a solution of* (12)–(16) *in the form*

$$f^+(x,\psi) = \psi + \epsilon(\psi - h)S(x) + \epsilon((\psi - h)\omega(x) + \theta^+(x,\psi)) \quad \text{for } 0 < \psi < h,$$
$$f^-(x,\psi) = \psi - h\epsilon(\psi + 1)S(x) + \epsilon(-h(\psi + 1)\omega(x) + \theta^-(x,\psi)) \quad \text{for } -1 < \psi < 0,$$

where

$$S(x) = \frac{\lambda_1 h}{h^2 - \rho}\operatorname{sech}^2\left(\left(\frac{\lambda_1}{\tau - \tau_0}\right)^{1/2} \epsilon^{1/2}\frac{x}{2}\right),$$

and $\omega(x) \in X^s$, $\theta \in B^s$ *with*

$$\|\omega(x)\|_{X^s} \leq K\epsilon \quad \text{and} \quad \|\theta(x,\psi)\|_{B^s} \leq K\epsilon$$

for $s \geq 4$, *where* K *is a constant independent of* ϵ. *The interface is given by*

$$\eta = \epsilon(-h)S(x) + O(\epsilon^2),$$

and $S(x)$ is indeed a first-order approximation to the exact solution.

5. Formal equations for the critical cases. In this section several critical cases are considered when the parameters in the problem are near some critical values so that (24) no longer holds. We shall only present the formal results here, and a rigorous justification of case 2 in the following will be given in §6. Cases 1 and 3 are similar to the results in [6], and their justifications will be deferred to subsequent investigation.

Case 1. τ is near $(1+\rho h)/3$ but h^2 is not near ρ. In (17)–(21), we let

$$\begin{aligned} &w = \epsilon^2 w_1 + \epsilon^3 w_2 + \epsilon^4 w_3 + \cdots, \\ &\nu = \lambda_0 + \epsilon^2 \lambda_1, \qquad \tau = \frac{1+\rho h}{3} + \epsilon \tau_1, \end{aligned}$$

and replace x with $\epsilon^{-1/2}x$. The equations for the first approximation are

$$\begin{aligned} &w_{1\psi\psi} = 0 \quad \text{in} \quad -1 < \psi < 0, \quad 0 < \psi < h, \\ &w_1^- = w_1^+, \qquad w_{1\psi}^- - \rho w_{1\psi}^+ - \lambda_0 w^- = 0 \quad \text{at } \psi = h, \\ &w = 0 \quad \text{at} \quad \psi = -1 \quad \text{and} \quad h. \end{aligned}$$

It is easily found that

$$\begin{aligned} &w_1^+ = \eta(x)(\psi - h), \qquad w_1^- = -h\eta(x)(\psi + 1), \\ &\lambda_0 = 1 + \frac{\rho}{h}, \end{aligned}$$

where $\eta(x)$ has to be determined.

The equations for the second-order approximation are

$$\begin{aligned} &w_{2\psi\psi} = -w_{1xx} \quad \text{in} \quad -1 < \psi < 0, \quad 0 < \psi < h, \\ &w_2^+ = w_2^-, w_{2\psi}^- - \rho w_{2\psi}^+ - \lambda_0 w_2^- = -\frac{(1+\rho h)w_{1xx}}{3} \quad \text{at } \psi = 0, \\ &w_2 = 0 \quad \text{at } \psi = -1 \quad \text{and} \quad h. \end{aligned}$$

Then

$$\begin{aligned} w_2^+ &= -\eta_{xx}\left(\frac{\psi^3}{6} - \frac{h\psi^2}{2} + \frac{h^2\psi}{3}\right) - a(x)\frac{\psi - h}{h}, \\ w_2^- &= h\eta_{xx}\left(\frac{\psi^3}{6} + \frac{\psi^2}{2} + \frac{\psi}{3}\right) + a(x)(\psi + 1), \end{aligned}$$

where $a(x)$ is an unknown function. Finally we proceed to the equations for the third approximation:

$$\begin{aligned} &w_{3\psi\psi} = -w_{2xx} \quad \text{in} \quad -1 < \psi < 0, \quad 0 < \psi < h, \\ &w_3^+ = w_3^-, w_{3\psi}^- - \rho w_{3\psi}^+ - \lambda_0 w_3^- = \lambda_1 w_1^- - \tau_1 w_{1xx} \\ &\qquad - \frac{(1+\rho h)w_{2xx}}{3} + \frac{3}{2}(h^2 - \rho)\eta^2 \quad \text{at } \psi = 0, \\ &w_3 = 0 \quad \text{at} \quad \psi = -1 \quad \text{and} \quad h. \end{aligned}$$

To solve these equations, the following solvability condition on $\eta(x)$ must hold

$$\lambda_1 \eta - \frac{3}{2}\left(h - \frac{\rho}{h}\right)\eta^2 - \tau_1 \eta_{xx} + \frac{(1+\rho h)\eta_{xxxx}}{45} = 0,$$

where $w_1^+ = \eta(x)(\psi - h)$, $w_1^- = -h\eta(x)(\psi + 1)$.

If $\lambda_1 > 0$, $\tau_1 > 0$, and

$$\lambda_1 = \frac{1620}{169}\left(\frac{\tau_1^2}{1+\rho h}\right),$$

This equation has a solution [6]

$$\eta(x) = 1225\tau_1^2\left(\left(h - \frac{\rho}{h}\right)169(1+\rho h)\right)^{-1}\operatorname{sech}^4(\alpha x),$$

$$\alpha = \frac{1}{2}\left(\frac{45\tau_1}{13(1+\rho h)}\right)^{1/2}.$$

Case 2. ρ is near h^2, but τ is not near $(1+\rho h)/3$. In this case we let

$$w = \epsilon w_1 + \epsilon^2 w_2 + \epsilon^3 w_3 + \epsilon^4 w_4 + \cdots,$$
$$\nu = \lambda_0 + \epsilon^2\lambda_1, \quad h^2 = \rho + \epsilon^2\sigma, \quad x \to \epsilon^{-1}x,$$

and substitute them for w, ν, and h^2 in (17)–(21). By the equations for the successive approximations we obtain

$$\left(\tau - \frac{1+h^3}{3}\right)\eta_{xx} - \left(\lambda + \frac{\sigma}{h}\right)\eta + 2h(1+h)\eta^3 = 0,$$

which also holds for $\sigma = 0$. A solitary wave solution of the above modified K-dV equation is given by

$$\eta(x) = \left(\frac{\lambda_1 + \sigma/h}{h(1+h)}\right)^{1/2}\operatorname{sech}\left(\left(\frac{\lambda_1+\sigma/h}{(\tau-(1+h^3))/3}\right)^{1/2}x\right),$$

where $\lambda_1 + (\sigma/h) > 0$ and $\tau > (1+h^3)/3$.

Case 3. τ is near $(1+\rho h)/3$ and h^2 is near ρ. Now we let

$$w = \epsilon^2 w_1 + \epsilon^3 w_2 + \epsilon^4 w_3 + \epsilon^5 w_4 + \cdots,$$
$$\rho = h^2 - \epsilon^4\sigma, \quad \nu = \lambda_0 + \epsilon^4\lambda_1, \quad \tau = \frac{1+h^3}{3} + \tau_1\epsilon^2,$$

in (17)–(21) and replace x by $\epsilon^{-1}x$. As before, from the equations for the successive approximations, we obtain

$$\frac{h}{45}(1+h^5)\eta_{xxxx} + \frac{h^2}{3}\eta\eta_{xx}(1-h^2) - \tau_1 h\eta_{xx} + \frac{2h^2}{3}(1-h^2)\eta_x^2$$
$$+ (\sigma + \lambda_1 h)\eta - 2h^2(1+h)\eta^3 = 0,$$

where $w_1^+ = \eta(x)(\psi - h)$, $w_1^- = -h\eta(x)(\psi + 1)$. Using the methods in [15], [16], we find that if $\tau_1 > 0$, $(\sigma + \lambda_1 h)(1+h^5) = 90(8-a)(10-a)^{-2}\tau_1^2$, then

$$\eta(x) = \frac{ab^2}{4}\operatorname{sech}^2\frac{abx}{2},$$

where

$$\begin{aligned} a &= \frac{-7 \pm (49 + 120t)^{1/2}}{t}, \\ t &= \frac{(2/5)(1+h)(1+h^5)}{h(1-h^2)^2}, \\ b &= \left(\frac{6\tau_1}{(1-h^2)h(10-a)} \right)^{1/2}, \\ \alpha^2 &= \frac{15(1-h^2)h}{1+h^5}. \end{aligned}$$

We note that here $8 - a > 0$ as can be easily shown, and $\tau_1 > 0$ and $\rho < 1$ implies $1 - h^2 > 0$ for sufficiently small ϵ.

6. Critical case for ρ near h^2. In this section, we shall justify the asymptotic method for the derivation of a solitary wave solution when ρ is near h^2 and $\tau > (1 + \rho h)/3$ but not near $(1 + \rho h)/3$. Since the existence proof is rather similar to the one presented in §3, many detailed derivations will be omitted.

First we rewrite (17) in the following form:

$$\begin{aligned} w_{\psi\psi} + w_{xx} &= (2w_x w_{\psi x} - w_x^2 w_{\psi\psi} + 2w_x w_\psi w_{x\psi})(1 + w_\psi)^{-2} \\ &\quad + w_{\psi\psi} w_\psi (2 + w_\psi)(1 + w_\psi)^{-2} = L_1(w), \end{aligned} \tag{65}$$

in $-1 < \psi < 0$ and $0 < \psi < h$. If we let

$$\begin{aligned} &\rho = h^2 - \sigma\epsilon^2, \quad \nu = \lambda_0 + \lambda_1\epsilon^2, \quad \tau > \frac{1 + h^3}{3}, \\ &\lambda_0 = 1 + h, \\ &w^+ = \epsilon(\eta(x)(\psi - h) + \theta^+(x, \psi)), \\ &w^- = \epsilon(\eta(x)(\psi + 1)(-h) + \theta^-(x, \psi)); \end{aligned}$$

in (65), (18)–(21), we obtain

$$\theta^+_{\psi\psi} + \theta^+_{xx} = -\eta_{xx}(\psi - h) + \frac{1}{\epsilon} L_1^+(w^+) = L_2^+(x, \psi), \qquad 0 < \psi < h, \tag{66}$$

$$\theta^-_{\psi\psi} + \theta^-_{xx} = h\eta_{xx}(\psi + 1) + \frac{1}{\epsilon} L_1^-(w^-) = L_2^-(x, \psi), \qquad -1 < \psi < 0, \tag{67}$$

at $\psi = 0$,

$$\theta^+(x, 0) = \theta^-(x, 0), \tag{68}$$

$$\begin{aligned} \theta^-_\psi - h^2\theta^+_\psi - \lambda_0\theta^- + \tau\theta_{xx} &= \tau h\eta_{xx} - (\lambda_1 h + \sigma)\eta\epsilon^2 + 2h^2(1 + h)\eta^3\epsilon^2 \\ &\quad + P_1^*(\eta, \theta) = G_2(x), \end{aligned} \tag{69}$$

where

$$\begin{aligned} P_1^*(\eta, \theta) = &-\sigma\epsilon^2\theta^+_\psi + \lambda_1\epsilon^2\theta^- - 2h^2(1 + h)\eta^3\epsilon^2 + \frac{3}{2}\epsilon(-h\eta + \theta^-_\psi)^2 \\ &- \frac{3}{2}(h^2 - \sigma\epsilon^2)\epsilon(\eta + \theta^+_\psi)^2 - 2\epsilon^2(-h\eta + \theta^-_\psi)^3 \\ &+ 2(h - \sigma\epsilon^2)\epsilon^2(\eta + \theta^+_\psi)^3 + (\epsilon^{-1}F_1^*(w) + 2\epsilon^2(-h\eta + \theta^-_\psi)^3 \\ &- 2(h^2 - \sigma\epsilon^2)\epsilon^2(\eta + \theta^+_\psi)^3). \end{aligned}$$

At $\psi = -1$ and $\psi = h_0$,

$$\theta = 0. \tag{70}$$

By the same reasoning as before, the solvability condition

$$\int_{-1}^{0} L_2^-(x,\psi)(-h)(\psi+1)d\psi + h^2 \int_0^h L_2^+(x,\psi)(\psi-h)d\psi + hG_2(x) = 0$$

implies

$$\begin{aligned}
&\left(\tau - \frac{1+h^3}{3}\right)\eta_{xx} - \left(\lambda_1 + \frac{\sigma}{h}\right)\epsilon^2\eta + 2h(1+h)\eta^3\epsilon^2 \\
&= \frac{1}{h^2}\left(-P_1^*(\eta,\theta) - \int_{-1}^{0} \frac{1}{\epsilon} L_1^-(w^-)(-h)(\psi+1)d\psi \right. \\
&\qquad \left. - h^2 \int_0^h \frac{1}{\epsilon} L_1^+(w^+)(\psi - h)d\psi \right) \\
&= \Psi_1^*(\eta,\theta).
\end{aligned} \tag{71}$$

If $\Psi_1^*(\eta,\theta) = 0$, then

$$T(\epsilon x) = \left(\frac{\lambda_1 + (\sigma/h)}{h(1+h)}\right)^{1/2} \operatorname{sech}\left(\frac{\lambda_1 + (\sigma/h)}{(\tau - (1+h^3)/3)^{1/2}}\epsilon x\right),$$

is a solution of (71), provided that $\lambda_1 + (\sigma/h) > 0$ and $\tau > (1+h^3)/3$ but not near $(1+h^3)/3$. Now let

$$\eta(x) = T(\epsilon x) + \omega(x).$$

Then by (71), we have

$$\begin{aligned}
&(\tau - \tau_0)\omega_{xx} - \left(\lambda_1 + \frac{\sigma}{h}\right)\epsilon^2\omega + 6h(1+h)\epsilon^2 T^2(\epsilon x)\omega(x) \\
&= -2h(1+h)\epsilon^2(3T(\epsilon x)\omega^2(x) + \omega^3(x)) + \Psi_1^*(T(\epsilon x) + \omega(x), \theta) \\
&= \Psi_1(\omega,\theta),
\end{aligned} \tag{72}$$

where $\tau_0 = (1+h^3)/3$. We need to prove that (66)–(70) and (72) possess a small solution (ω,θ).

We define X^s, Y^s, and B^s as before except that we need $f(x)\cosh(\mu\epsilon x) \in H^s(\mathbf{R})$ in the definition of X^s instead of $f(x)\cosh(\mu\epsilon^{1/2}x) \in H^s(\mathbf{R})$. Since $\varphi(x) = T(\epsilon x)$ is a solution of (72) with $\Psi_1(\omega,\theta) = 0$, it is not difficult to construct another linearly independent solution $\psi(x)$ with $\psi'(0) = 0$. Also we assume that Wronskian $(\varphi,\psi) = -1$. Therefore, if we let all functions be even in x, the even solution in x of (72) is

$$\begin{aligned}
\omega(x) &= \varphi(x)\int_0^x \psi(s)\Psi_1(\omega,\theta)ds + \psi(x)\int_x^{+\infty} \varphi(s)\Psi_1(\omega,\theta)ds \\
&= \int_0^{+\infty} k_1^*(x,s)\Psi_1(\omega,\theta)ds \quad \text{for } x \geq 0,
\end{aligned}$$

and

$$\omega(x) = \int_0^{+\infty} k_1^*(-x,s)\Psi_1(\omega,\theta)ds \quad \text{for } x < 0 .$$

We write $\omega(x) = \mathcal{L}^{-1}(\Psi_1(\omega,\theta))$. Similar to Lemmas 1 and 2, we have the following.

LEMMA 6. *If $f(x) \in X^s$, $s \geq 0$ and μ is small, then*

$$\|\mathcal{L}^{-1}(f(x))\|_{X^{s+2}} \leq K\epsilon^{-2}\|f(x)\|_{X^s}.$$

LEMMA 7. *If $u_1 \in H^s(\mathbf{R})$, $u_2 \in X^s$, $s \geq 1$, then*

$$\|u_1u_2\|_{X^s} \leq K\|u_1\|_{H^s}\|u_2\|_{X^s}.$$

The proofs are the same as before and omitted. Note that the functions in X^s and Y^s here have the different decay rate with the functions of these spaces in §4 when $x \to \infty$. By a refinement of the proof of Lemma 6, we have the following.

LEMMA 8. *If $f(x) \in X^s$, $s \geq 0$ and μ is small, then*

$$\left\|\frac{d}{dx}(\mathcal{L}^{-1}(f))\right\|_{X^s} \leq K\epsilon^{-1}\|f(x)\|_{X^s} .$$

The proof is omitted. Note that to prove Lemma 8, differentiating $\varphi(x)$ and $\psi(x)$ once will introduce an ϵ-factor.

Now we can write $\Psi_1(\omega,\theta)$ as follows:

$$\begin{aligned}\Psi_1(\omega,\theta) &= \frac{\epsilon}{6}h(1-h^2)(\omega+T(\epsilon x))_x^2 + \Psi_2(\omega,\theta)\\ &= I_1 + \Psi_2(\omega,\theta) .\end{aligned}$$

If $\omega,\theta \in B^s$ and $\|\omega\|_{X^s} \leq K\epsilon^{1/2}$, $\|\theta\|_{B^s} \leq K\epsilon^2$ with $s \geq 3$, then $\Psi_1(\omega,\theta) \in X^{s-2}$ and $\|\Psi_2(\omega,\theta)\|_{X^{s-2}} \leq K\epsilon^3$ by checking the terms in $\Psi_2(\omega,\theta)$ and $\Psi_1(\omega,\theta)$. Now if $\|\omega_x\|_{X^{s-2}} \leq K\epsilon^{3/2}$, then $\|I_1\|_{X^{s-2}} \leq K\epsilon^3$. Therefore, by Lemmas 6, 7, and 8, we have the following.

THEOREM 6. *If $\omega,\theta \in B^s$, $\|\omega\|_{X^s} \leq K\epsilon^{1/2}$, $\|\theta\|_{B^s} \leq K\epsilon^2$, and $\|\omega_x\|_{X^{s-2}} \leq K\epsilon^{3/2}$, then*

$$\|\mathcal{L}^{-1}\Psi_1(\omega,\theta)\|_{X^s} \leq K\epsilon$$

and

$$\left\|\frac{d}{dx}\mathcal{L}^{-1}\Psi_1(\omega,\theta)\right\|_{X^{s-2}} \leq K\epsilon^2.$$

For the partial differential equations (66)–(70), we have exactly the same theorems as Theorems 2 and 3. Let

$$\begin{aligned}L_2(x,\psi) = L^*(\omega,\theta) &= M_1(\omega) + \tilde{L}_2(\omega,\theta),\\ G_2(x) = G^*(\omega,\theta) &= M_2(\omega) + \tilde{G}_2(\omega,\theta),\end{aligned}$$

where $M_1(\omega)$ and $M_2(\omega)$ have the terms with $\omega_{xx}, \epsilon^2\omega$, and $\epsilon^2T\omega$ only in L_2 and G_2, respectively, and $\tilde{L}_2$ and $\tilde{G}_2$ are the remainders. Define

$$\begin{aligned}L(\omega,\theta) &= M_1(\mathcal{L}^{-1}(\Psi_1(\omega,\theta))) + \tilde{L}_2(\omega,\theta),\\ G(\omega,\theta) &= M_2(\mathcal{L}^{-1}(\Psi_1(\omega,\theta))) + \tilde{G}_2(\omega,\theta),\end{aligned}$$

where $\mathcal{L}(\omega) = \Psi_1(\omega, \theta)$ is the same as (72). Then $(L(\omega,\theta), G(\omega,\theta)) \in H$ for all ω and θ by a similar proof as this for F and G in §4.3. If we use the same notation $\mathcal{P}^{-1}$ as in (58), then we have the following theorem.

THEOREM 7. *If $\theta(x,\psi) \in B^{s+2}$, $\omega \in X^{s+2}$ with $\|\theta\|_{B^{s+2}} \leq K\epsilon$, $\|\omega\|_{X^{s+2}} \leq K\epsilon^{1/2}$, then the solution of* (66)–(70) *with* (L_2, G_2) *replaced by* (L, G) *exists and satisfies*

$$\|\mathcal{P}^{-1}(L,G)\|_{B^{s+2}} \leq K\epsilon^2(\|\theta\|_{B^{s+2}} + \|\omega\|_{X^{s+2}}).$$

Now we need to solve

$$\theta = \mathcal{P}^{-1}(L_2, G_2) = \mathcal{P}^{-1}(L,G) = \mathcal{T}_1(\theta,\omega) \tag{73}$$

and

$$\omega = \mathcal{L}^{-1}(\Psi_1(\omega,\theta)) = \mathcal{R}_1(\theta,\omega) \tag{74}$$

when ϵ is small. Then we substitute (73) in (74) once to obtain

$$\omega = \mathcal{R}_1(\mathcal{T}_1(\theta,\omega),\omega) = \mathcal{R}_2(\theta,\omega), \tag{75}$$

and another substitution yields

$$\omega = \mathcal{R}_2(\mathcal{T}_1(\theta,\omega),\omega) = \mathcal{T}_2(\theta,\omega). \tag{76}$$

Obviously (76) is equivalent to (74). Let us define a closed convex set in the Banach space $B^{s+2} \times X^{s+2}$ for $s \geq 2$,

$$\begin{aligned}\mathcal{S}_b = \big\{Z = (\theta,\omega) \in B^{s+2} \times X^{s+2} \mid |||Z||| &= \|\omega\|_{X^{s+2}} \\ + \|\theta\|_{X^{s+2}} < +\infty, \|\omega\|_{X^{s+2}} &+ \epsilon^{-1}\|\theta\|_{B^{s+2}} + \epsilon^{-1}\|\omega_x\|_{X^s} \leq b\epsilon^{1/2}\big\}.\end{aligned}$$

Let

$$\mathcal{T}(\theta,\omega) = (\mathcal{T}_1(\theta,\omega), \mathcal{T}_2(\theta,\omega)).$$

Then by Theorems 6 and 7, we have that $\mathcal{T}$ maps $\mathcal{S}_b$ into itself if ϵ is small. By checking L and G term by term and using the proof of Theorem 7, we have that if $Z^{(1)}$, $Z^{(2)} \in \mathcal{S}_b$,

$$\begin{aligned}&\|\mathcal{T}_1(\theta^{(1)},\omega^{(1)}) - \mathcal{T}_1(\theta^{(2)},\omega^{(2)})\|_{B^{s+2}} \\ &\qquad \leq K\epsilon(\|\theta^{(1)} - \theta^{(2)}\|_{B^{s+2}} + \epsilon^{1/2}\|\omega^{(1)} - \omega^{(2)}\|_{X^{s+2}}).\end{aligned} \tag{77}$$

Also by checking the terms in Ψ_1 and using Theorem 6, we have that if $Z^{(1)}$, $Z^{(2)} \in \mathcal{S}_b$, then

$$\|\mathcal{R}_2(\theta^{(1)},\omega^{(1)}) - \mathcal{R}_2(\theta^{(2)},\omega^{(2)})\|_{X^{s+2}} \leq K(\|\theta^{(1)} - \theta^{(2)}\|_{B^{s+2}} + \epsilon^{1/2}\|\omega^{(1)} - \omega^{(2)}\|_{X^{s+2}}).$$

By the definition of $\mathcal{T}_2$ and (77), we have

$$\|\mathcal{T}_2(\theta^{(1)},\omega^{(1)}) - \mathcal{T}_2(\theta^{(2)},\omega^{(2)})\|_{X^{s+2}} \leq K\epsilon^{1/2}(|||Z^{(1)} - Z^{(2)}|||).$$

Therefore if $K\epsilon^{1/2} \leq \frac{1}{2}$, then $\mathcal{T}$ is a contraction and possesses a fixed point in $\mathcal{S}_b$. So (73) and (74) have a solution in $\mathcal{S}_b$. Finally we summarize our results in the following.

THEOREM 8. *If* $\nu = (1+h)+\lambda_1\epsilon^2$, $\rho = h^2-\sigma\epsilon^2$, $\tau > (1+h^3)/3$ *with* $\lambda_1+\sigma/h > 0$, *then for small* $\epsilon > 0$, *there exists a solution of* (12)–(16) *in the form*

$$\begin{aligned}
f^+(x,\psi) &= \psi + \epsilon(\psi - h)T(\epsilon x) + \epsilon((\psi - h)\omega(x) + \theta^+(x,\psi)), \qquad 0 < \psi < h,\\
f^-(x,\psi) &= \psi - h(\psi + 1)T(\epsilon x)\epsilon + \epsilon(-h(\psi + 1)\omega(x) + \theta^-(x,\psi)), \qquad -1 < \psi < 0,
\end{aligned}$$

where

$$T(\epsilon x) = \left(\left(\frac{\lambda_1 + \sigma/h}{h(1+h)}\right)^{1/2} \operatorname{sech}\left(\frac{\lambda_1 + (\sigma/h)}{\tau - (1+h^3)/3}\right)^{1/2} \epsilon x\right),$$

and

$$\|\omega(x)\|_{X^s} \le K\epsilon, \qquad \|\theta(x,\psi)\|_{B^s} \le K\epsilon^2 \quad \text{and} \quad \|\omega_x\|_{X^{s-2}} \le K\epsilon^2,$$

for $s \ge 4$.

Appendix. Let

$$F(x) = -\Gamma(x) = \lambda_0 + \tau x - \sqrt{x}\left(\frac{\cosh\sqrt{x}}{\sinh\sqrt{x}} + \rho\frac{\cosh(\sqrt{x}h)}{\sinh(\sqrt{x}h)}\right). \tag{A.1}$$

Obviously as $x \to 0$, $F(x) \to 0$. We need to prove that $F(x) \ne 0$ for all x if $\tau > \tau_0 = (1+\rho h)/3$. We differentiate (A.1) to get

$$\begin{aligned}
F'(x) =& \tau - \frac{1}{4}\frac{\sinh(2x^{1/2}) - 2x^{1/2}}{x^{1/2}\sinh^2 x^{1/2}}\\
&- \frac{(\rho h/4)\sinh(2hx^{1/2}) - 2hx^{1/2}}{hx^{1/2}\sinh^2(hx^{1/2})},
\end{aligned}$$

and as $x \to 0$, $F'(x) \to \tau - (1+\rho h)/3 = \tau - \tau_0$. Now let $I(y) = (\sinh(2y) - 2y)/(y\sinh^2 y)$. Then

$$\begin{aligned}
F''(x) &= -\frac{1}{8}x^{1/2}I_y(x^{1/2}) - \frac{\rho}{8}x^{1/2}I_y(hx^{1/2}),\\
I_y(y) &= (2y\cosh(2y)\sinh^2 y - \sinh^2 y\sinh 2y - y\sinh^2 2y + 2y^2\sinh^2 y)\\
&\quad \times (y^2\sinh^4 y)^{-1} = \frac{I_1}{I_2}.
\end{aligned}$$

Now we show that for $y > 0$, $I_1 < 0$. By using some elementary identities, we have

$$\begin{aligned}
I_1(y) &= -2y\sinh^2 y - \sinh(2y)\sinh^2 y + 2y^2\sinh 2y\\
&= \sinh y\left(-2y\sinh y - \frac{1}{2}\cosh 3y + \frac{1}{2}\cosh y + 4y^2\cosh y\right)\\
&= \frac{1}{2}\sinh y(-4y\sinh y - \cosh 3y + \cosh y + 8y^2\cosh y)\\
&= \frac{1}{2}\sinh y\left(\sum_{n=0}^{\infty}(-4(2n+2) - 3^{2n+2}\right.\\
&\qquad \left. + 1 + 8(2n+2)(2n+1))y^{2n+2}/(2n+2)!\right).
\end{aligned}$$

It is easy to check that when $y \geq 3$,

$$3^y > y^2 + 3y + 1.$$

So $3^{2n} > (2n+2)(2n+1)$ for $n \geq 2$ and $3^{2n+2} > 8(2n+2)(2n+1)$. But for $n = 0$ and $n = 1$,

$$-4(2n+2) - 3^{2n+2} + 1 + 8(2n+2)(2n+1) = 0.$$

Therefore, $I_1(y) < 0$ for $y > 0$. So $F''(x) > 0$ for $x > 0$ and $F''(x) \to c_1 > 0$ as $x \to 0$ for some constant c_1. $F'(x)$ is monotonically increase for $x \geq 0$. If $\tau > \tau_0$, then $F'(x) > 0$, which means $F(x)$ is monotonically increase and so $F(x) > 0$ for all $x > 0$.

If $\sigma < 0$ by (45)–(47) σ should satisfy the equation

(A.2)

$$(\lambda_0 + \tau\sigma) \sin\sqrt{-\sigma}h \sin\sqrt{-\sigma} = \sqrt{-\sigma} \sin\sqrt{-\sigma}h \cos\sqrt{-\sigma} + \rho\sqrt{-\sigma} \sin\sqrt{-\sigma} \cos\sqrt{-\sigma}h.$$

If σ is large, then (A.2) becomes

$$\tau \sin(\sqrt{-\sigma}h) \sin\sqrt{-\sigma} = O\left(\frac{1}{(-\sigma)^{1/2}}\right).$$

Therefore, for larger σ, there is always a solution near each of the solutions of

$$\sin((-\sigma)^{1/2}h) \sin(-\sigma)^{1/2} = 0,$$

which means that the roots of (A.2) are infinite and σ_n has order of n^2 when n is large.

Remark. When $\tau < \tau_0$, we know that $F(x)$ in (A.1) is zero at $x = 0$, $F''(x) > 0$ for $x \geq 0$, $F'(0) = \tau - \tau_0 < 0$ and $F(x) \to +\infty$ as $x \to +\infty$. From these facts, it is easy to see that $F(x)$ has one and only one positive number $x_0 > 0$ so that $F(x_0) = 0$ and $F'(x_0) \neq 0$ since $F(x)$ is a strictly convex function for $x \geq 0$.

REFERENCES

[1] D. J. Korteweg and G. DeVries, *On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves*, Phil. Mag., 39 (1895), pp. 422–443.

[2] K. O. Friedrichs and D. Hyers, *The existence of solitary waves*, Comm. Pure Appl. Math., 3 (1954), pp. 517–550.

[3] J. K. Hunter and J.-M. Vanden-Broeck,*Solitary and periodic gravity-capillary waves of finite amplitude*, J. Fluid Mech., 134 (1983), pp. 205–219.

[4] C. J. Amick and K. Kirchgässner, *A theory of solitary water-waves in the presence of surface tension*, Arch. Rational Mech. Anal., 105 (1989), pp. 1–49.

[5] R. L. Sachs, *On the existence of small amplitude waves with strong surface tension*, J. Differential Equations, 90 (1991), pp. 31–51.

[6] S. M. Sun and M. C. Shen, *A new solitary wave solution for water waves with surface tension*, Ann. Mat. Pure Appl., 162 (1992), pp. 179–214.

[7] J. K. Hunter and J. Scheurle, *Existence of perturbed solitary wave solutions to a model equation for water waves*, Physica D, 32 (1988), pp. 253–268.

[8] J. T. Beale, *Exact solitary water waves with capillary ripples at infinity*, Comm. Pure Appl. Math., 44 (1991), 211–257.

[9] S. M. Sun, *Existence of a generalized solitary wave solution for water with positive Bond number less than* 1/3, J. Math. Anal. Appl., 156 (1991), pp. 471–504.

[10] A. S. Peters and J. J. Stoker, *Solitary waves in liquids having non-constant density*, Comm. Pure Appl. Math., 13 (1960), pp. 115–164.

[11] C. J. Amick and R. E. L. Turner, *Small internal waves in two-fluid systems*, Arch. Rational Mech. Anal., 108 (1989), pp. 111–139.

[12] R. E. L. Turner and J. M. Vanden-Broeck, *Broadening of interfacial solitary waves*, Phys. Fluids, 31 (1988), pp. 2486–2490.

[13] J. Bona and R. L. Sachs, *The existence of internal solitary waves in a two-fluid system near the KdV limit*, Geophys. Astrophys. Fluid Dynamics, 48 (1989), pp. 25–51.

[14] A. M. Ter-Krikorov, *Théorie exact des ondes longues stationnaires dans un liquide hétéogéne*, J. de Mécanique, 3 (1963), pp. 351–375.

[15] R. Hirota, *Direct method of finding exact solutions of nonlinear evolution equations*, in Lecture Notes in Math., 515, R. M. Miura, ed., Springer-Verlag, New York, 1976, pp. 40–68.

[16] S. Melkonian, *Nonlinear waves on thin films*, in Continuum Mechanics and Its Applications, G. A. C. Grapham and S. K. Malik, eds., Hemisphere, New York, 1989, pp. 411–423.

SIAM J. MATH. ANAL.
Vol. 24, No. 4, pp. 892–910, July 1993

004

THE LINEAR SHALLOW WATER THEORY: A MATHEMATICAL JUSTIFICATION*

JAMES A. DONALDSON[†][‡] AND DANIEL A. WILLIAMS[†][§]

Abstract. The authors provide in the space of square integrable measurable functions a mathematical justification for the "shallow water" theory for time-dependent two-dimensional flows of an inviscid, irrotational, incompressible fluid moving under the influence of gravity. A by-product of this investigation is a new derivation of the "shallow water" equations.

Key words. shallow water equations, elliptic boundary-value problems, Cauchy problems, generalized transforms

AMS subject classifications. 35J25, 35L15, 35Q35, 76B15, 76D30

1. Introduction. We are concerned with a problem from the theory of water waves [15] which arises when one makes two basic assumptions: (i) the depth of the water is small (relative to some other quantity associated with the problem) and (ii) the amplitude of surface waves is small (relative to their wavelength). This leads to the following mathematical model, when we assume also that the fluid is incompressible, inviscid, and irrotational. Let the positive direction of the Y-axis be vertically upward; let the X-axis be at the mean height of the surface of the water at time $t = 0$; and let $B(X)$ be a function contained in $C^3(\mathbb{R})$. We are required to find a function $\Phi(X, Y, t)$ such that Φ satisfies the boundary-value problem

$$\Phi_{XX} + \Phi_{YY} = 0 \quad \text{in } \Omega_B \times \mathbb{R}^+; \tag{1.1}$$

$$\frac{\partial \Phi}{\partial n} = 0 \quad \text{on } \Gamma_B; \tag{1.2}$$

$$\Phi_Y = \eta_t, \qquad \Phi_t + \eta = 0 \quad \text{on } \Gamma \times \mathbb{R}^+; \tag{1.3a}$$

or equivalently,

$$\Phi_Y + \Phi_{tt} = 0 \quad \text{on } \Gamma \times \mathbb{R}^+; \tag{1.3b}$$

$$\Phi(X, 0, 0) = F_0(X) \quad \text{on } \Gamma; \tag{1.4a}$$

$$\Phi_t(X, 0, 0) = F_1(X) \quad \text{on } \Gamma, \tag{1.4b}$$

and where $Y = \eta(X, t)$ is the equation of the free surface,

$$\begin{aligned} &\Gamma = \{(X, 0) : X \in \mathbb{R}\}, \qquad \Gamma_B = \{(X, Y) : Y = -\epsilon B(X)\}, \\ &\Omega_B = \{(X, Y) : -\epsilon B(X) < Y < 0, X \in \mathbb{R}\}, \quad \text{and} \\ &\mathbb{R}^+ = \{r : r > 0\}. \end{aligned}$$

*Received by the editors October 20, 1990; accepted for publication (in revised form) October 19, 1992.

[†] Mathematics Department, Howard University, Washington, DC 20059.

[‡] This author's research was supported in part by National Science Foundation grant R11-891-2667.

[§] This author's research was supported in part by Air Force Office of Scientific Research grant R1NM062.

The initial conditions (1.4a,b) result from specifying the initial surface elevation $\eta(X,0)$ and the initial velocity potential $\Phi(X,Y,0)$. We assume (i) there exist positive constants b_1, b_2, and b_3 such that $b_1 \leq B(X) \leq b_2$ and $B'(X) < b_3$, (ii) the parameter ϵ is a small positive number, and (iii) the density of the fluid and the acceleration due to gravity are constants which we equate to one. Furthermore, when we assume that B has bounded derivatives through order n in $\mathbb{R}$, B_n will denote $\sup_{x\in\mathbb{R}}\{\sum_{k=0}^{n} |B^{(k)}(x)|\}$.

For ϵ fixed (but not necessarily small) and F_0 and F_1 sufficiently smooth, Friedman and Shinbrot [4] and Garipov [6] have proved the existence of smooth solutions of (1.1)–(1.4) satisfying all conditions classically. Furthermore, they defined in $L^2(\Gamma)$ a self-adjoint (nonlocal) operator K with domain $H^1(\Gamma)$ related to (1.1)–(1.4) by the expression

$$\Phi_y(X,0,t) = K\Phi(X,0,t) = KW(X,t), \tag{1.5}$$

where $W(X,t)$ is the restriction to Γ of the unique solution $\Phi(X,Y,t)$ of (1.1)–(1.4) in $H^1(\Omega)$.

In the linear theory of "shallow water" (ϵ small), the hyperbolic initial-value problem

$$W^0_{tt} - \epsilon(BW^0_X)_X = 0, \quad W^0(X,0) = F_0(X), \quad W^0_t(X,0) = F_1(X) \tag{1.6}$$

plays an essential role. The partial differential equation in this problem is called the "shallow water" equation. Naturally, the replacement of the elliptic boundary value problem (1.1)–(1.4) by the hyperbolic initial-value problem in (1.6) in the "shallow water" theory requires justification.

Friedrichs [5] gave a derivation of the shallow water theory by a formal perturbation procedure in powers of a small parameter $\sigma = \kappa h$, where κ is the curvature of the surface at $t = 0$ and h is the depth of the water. Friedrichs' requirement that σ be small for the water to be considered "shallow" is equivalent to the small depth assumption made in (1.1)–(1.4a) since for this case σ is small whenever ϵ is. (Note: $0 < \kappa\epsilon b_1 \leq \sigma \leq \kappa\epsilon b_2$ and we assume the curvature is bounded.) Friedrichs stated that a rigorous mathematical justification of the "shallow water" theory requires a proof that the perturbation series is convergent, or asymptotically valid. For the case of simple harmonic motion of two-dimensional linear flows, this justification was provided by Shinbrot [13].

For the case of two-dimensional nonlinear flows many authors have contributed. In the space of analytic functions mathematical justifications for the theory have been given by Ovsjannikov [11] valid for periodic solutions, and by Kano and Nishida [8] valid for sufficiently small time t. In Sobolev space Nalimov [10] proved unique solvability when the region of the fluid has infinite depth, and gave a justification for the linear theory, Yosihara [16] proved unique solvability when the region of the fluid has finite depth and the bottom is almost horizontal, and Craig [1] proved unique solvability and provided mathematical justifications for the Boussinesq equations and the Korteweg–deVries equations when the region of the fluid has finite depth and the bottom is horizontal.

In a Sobolev space we provide a mathematical justification for the "shallow water" theory in the case of time-dependent two-dimensional linear flows where the region of the fluid has finite depth and the bottom is neither horizontal nor almost horizontal.

We recall from (1.3a) that

$$\eta(X,t) = -\Phi_t(X,0,t) = -W_t(X,t)$$

and define a function η^0 by

$$\eta^0(X,t) = -W_t^0(X,t),$$

where $W_t^0(X,t)$ is the solution of the initial-value problem in (1.6). A fundamental problem in the "shallow water" theory is to determine in terms of ϵ a bound for the error which results when η^0 is used to approximate η. (This, of course, would provide the mathematical justification.) It is the aim of this paper to obtain such a bound valid for all t in $[0, S]$, where S is an arbitrary fixed positive number. We prove the following.

THEOREM 1.1. *Let S be any positive number, and let $F_0 \in H^2(\Gamma)$ and $F_1 \in H^1(\Gamma)$. Then there exits a constant C, depending upon S, B_3, $\|F_0\|_{H^2(\Gamma)}$, and $\|F_1\|_{H^1(\Gamma)}$ such that*

$$\|\eta(X,t) - \eta^0(X,t)\|_{L^2(\Gamma)} \le C\epsilon^{1/2}$$

for $t \in [0, S]$.

A brief outline of the paper follows. In §2 we give definitions and notation, introduce a generalization of the Fourier transform and study some of its properties, and state an inequality arising from an abstract Cauchy problem.

In §3 the region Ω_B is mapped into a horizontal strip Ω. Under this mapping the Laplacian operator is sent into an elliptic operator L with variable coefficients, and the problem (1.1)–(1.4) is transformed into an initial-boundary value problem that is rewritten, after performing a separation of variables, as a well-posed time-independent elliptic boundary-value problem and an initial-value problem. Introduced then is a boundary-value problem involving an elliptic operator L_1 related to L. The boundary-value problems for L and L_1 play a fundamental role in defining on $L^2(\Gamma)$ two invertible, positive, self-adjoint operators T and T_1. The operator T is related to the previously mentioned operator K by $K = B(T^{-1} - \alpha I)$, and the operator T_1 has the virtue of having an explicit representation. The proof of a key inequality employed in the demonstration of Theorem 1.1 is given.

In §4 an apparently new derivation of the "shallow water" equations is given. This derivation depends heavily upon an explicit representation obtained for an operator K_1 defined by $K_1 = B(T_1^{-1} - \alpha I)$ and an implicit representation for the aforementioned operator K. Introduced also is another operator K_0, with domain $H^2(\Gamma)$, and an associated abstract Cauchy problem for which certain estimates for its solution and derivatives of its solutions are obtained.

In §5 the estimates obtained in §4 and the results of §3 are used to establish the estimate needed to provide a rigorous justification of the "shallow water" theory.

2. Preliminary results and definitions.

We define $H^s(\Gamma)$ to be the collection of tempered distributions u such that $(1 + |\lambda|^2)^{s/2}\hat{u}(\lambda)$ is in $L^2(\Gamma)$, normed by

$$\|u\|_{H^s(\Gamma)} = \left\{ \int_\Gamma (1 + |\lambda|^2)^s |\hat{u}(\lambda)|^2 d\lambda \right\}^{1/2},$$

where $\hat{u}$ is the Fourier transform of u. We denote by $H^s(\Omega)$ the collection of all functions which, together with their first s distributional derivatives, are in $L^2(\Omega)$.

This space is normed by

$$\|u\|_{H^s(\Omega)} = \left\{ \sum_{|\alpha|\le s} \int_\Omega |D^\alpha u|^2 dx dy \right\}^{1/2} .$$

For $f \in L^2(\mathbb{R})$ we define a transformation, a slight generalization of the Fourier transformation, by the formula

$$\tilde{f}(\lambda) = \mathcal{F}_B[f](\lambda) = \int_{\mathbb{R}} f(x) e^{i\lambda H(x)} dH,$$

where

$$H(x) = \int_0^x \frac{ds}{B(s)}.$$

We shall call $\tilde{f}$ the B-transform of f. One verifies easily the following properties of this transform.

LEMMA 2.1. *If $f \in C_0^\infty(\mathbb{R})$ and $B \in C^{m-1}(\mathbb{R})$, then*

$$\mathcal{F}_B\left[\left(B\frac{\partial}{\partial x}\right)^n f\right] = (-i\lambda)^n \mathcal{F}_B[f](\lambda), \qquad 0 \le n \le m.$$

LEMMA 2.2. *Let $f \in L^2(\mathbb{R})$. Then there exist constants C_1 and C_2 such that*

$$C_1\|f\|_{L^2(\mathbb{R})} \le \|\tilde{f}\|_{L^2(\mathbb{R})} \le C_2\|f\|_{L^2(\mathbb{R})}.$$

LEMMA 2.3. *Let B have bounded derivatives through order $n-1$. Then there exist constants M_1 and M_2 such that*

$$\|\lambda^n \tilde{f}(\lambda)\|_{L^2(\mathbb{R})} \le M_1 \sum_{1\le j\le n} \|\lambda^j \hat{f}(\lambda)\|_{L^2(\mathbb{R})},$$

and

$$\|\lambda^n \hat{f}(\lambda)\|_{L^2(\mathbb{R})} \le M_2 \sum_{1\le j\le n} \|\lambda^j \tilde{f}(\lambda)\|_{L^2(\mathbb{R})}.$$

LEMMA 2.4. *The norm $\| \ \|_{B^k}$, defined by*

$$\|f\|_{B^k}^2 = \int_{\mathbb{R}} (1+\lambda^2)^k |\tilde{f}(\lambda)|^2 d\lambda,$$

is equivalent to the norm $\|f\|_{H^k(\mathbb{R})}$.

LEMMA 2.5. *Let U satisfy the abstract Cauchy problem*

$$\begin{aligned} &U''(t) + LU(t) = G(t), \qquad t > 0, \\ &U(0) = U'(0) = 0, \end{aligned}$$

where L is a self-adjoint positive operator defined in a Hilbert space X, and $G(t)$ is an X-valued strongly continuous function of t, and let S be a fixed, but arbitrarily selected positive number. Then

$$\|U'(t)\|_X \le \int_0^S \|G(s)\|_X ds \quad \textit{for all } \ t \in [0, S].$$

3. Two associated elliptic boundary-value problems. We introduce new independent variables x and y and dependent variable ϕ, defined by

$$x = X, \qquad y = \frac{Y}{\epsilon B(X)} \quad \text{and} \quad \phi(x,y,t) = \Phi(X,Y,t).$$

Introduction of these variables transforms the domain Ω_B into a horizontal strip Ω, and sends (1.1)–(1.4) into the problem

$$L\phi = 0 \quad \text{in } \Omega \times \mathbb{R}^+, \tag{3.1}$$

$$M\phi = 0 \quad \text{on } \Gamma_b \times \mathbb{R}^+, \tag{3.2}$$

$$\phi_y + \epsilon B\phi_{tt} = 0 \quad \text{on } \Gamma \times \mathbb{R}^+, \tag{3.3}$$

$$\phi(x,0,0) = F_0(x), \qquad \phi_t(x,0,0) = F_1(x) \quad \text{on } \Gamma, \tag{3.4}$$

where

$$\begin{aligned} L\phi &= \phi_{yy} + \epsilon^2[B(B\phi_x)_x + B_x^2 y^2\phi_{yy} - 2BB_x y\phi_{xy} + (2B_x^2 - BB_{xx})y\phi_y], \\ M\phi &= \phi_y + \epsilon^2[B_x^2\phi_y + BB_x\phi_x], \\ \Omega &= \{(x,y) : -1 < y < 0, x \in \mathbb{R}\}, \qquad \Gamma_b = \{(x,-1) : x \in \mathbb{R}\}, \end{aligned}$$

and

$$\Gamma = \{(x,0) : x \in \mathbb{R}\}.$$

Let $\alpha > 0$, and let $f(x,t) \in C_0^\infty(\Gamma)$ for each fixed t. We rewrite the condition (3.3) in the form

$$\{\phi_y + (\alpha B\phi - Bf)\} + \{-(\alpha B\phi - Bf) + \epsilon B\phi_{tt}\} = 0$$

and note that this permits us to obtain the solution of (3.1)–(3.4) from a boundary-value problem and an initial-value problem. More specifically, we associate with (3.1)–(3.4) the boundary-value problem

$$L\psi = 0 \quad \text{in } \Omega, \tag{3.5}$$

$$M\psi = 0 \quad \text{on } \Gamma_b, \tag{3.6}$$

$$\psi_y + \alpha B\psi = Bf(x,t) \quad \text{on } \Gamma, \quad f \in C_0^\infty(\Gamma \times \{t\}), \tag{3.7}$$

$$\psi \in H^1(\Omega), \tag{3.8}$$

and observe that its solution ψ satisfies (3.1)–(3.4) if f is chosen so that the initial-value problem

$$-\epsilon\psi_{tt} + \alpha\psi = f(x,t) \quad \text{on } \Gamma \times \{0\} \times \mathbb{R}^+, \tag{3.9}$$

$$\psi(x,0,0) = F_0(x), \qquad \psi_t(x,0,0) = F_1(x) \quad \text{on } \Gamma, \tag{3.10}$$

is satisfied. Essentially what has happened here is a separation of variables.

THEOREM 3.1. *For each fixed $\epsilon > 0$ the boundary-value problem* (3.5)–(3.8) *has a unique solution.*

Proof. Rewriting (3.5)–(3.8) in the Ψ, X, Y variables and employing the invertible transformation

$$(x, y, \psi) \to (X, Y, \Psi),$$

defined by

$$X = x, \ Y = -\epsilon B(x) y, \qquad \Psi(X, Y) = \psi(x, y),$$

we obtain an elliptic boundary-value problem for which existence and uniqueness results have been given by Friedman and Shinbrot [3] and Garipov [6]. □

We consider also the related boundary-value problem

$$(3.11) \qquad L_1 \psi^1 = \psi^1_{yy} + \epsilon^2 B(B\psi^1_x)_x = 0 \quad \text{in } \Omega,$$

$$(3.12) \qquad M_1 \psi^1 = \psi^1_y = 0 \quad \text{on } \Gamma_b,$$

$$(3.13) \qquad \psi^1_y + \alpha B \psi^1 = B f_1(x, t) \quad \text{on } \Gamma, \qquad f_1 \in C_0^\infty(\Gamma),$$

$$(3.14) \qquad \psi^1 \in H^1(\Omega).$$

THEOREM 3.2. *There exists one and only one solution of* (3.11)–(3.14).

For $f \in C_0^\infty(\Gamma)$, let $\psi(x, y; f)$ and $\psi^1(x, y; f)$ be the unique solutions of (3.5)–(3.8) and (3.11)–(3.14) corresponding to boundary datum f prescribed in (3.7) and (3.13). We define operators $\bar{T}$ and $\bar{T}_1$ on $C_0^\infty(\Gamma)$ by the formulas

$$(3.15) \qquad \bar{T} f = \psi(x, 0; f)$$

and

$$(3.16) \qquad \bar{T}_1 f = \psi^1(x, 0; f).$$

THEOREM 3.3. *The operator $\bar{T}$ defined by* (3.15) *has an extension to a bounded linear operator T on $L^2(\Gamma)$. The operator T is positive, self-adjoint, and has no nontrivial null-space.*

THEOREM 3.4. *The same statement as Theorem* 3.3, *but with $\bar{T}_1$ and T_1 replacing $\bar{T}$ and T, respectively.*

The proofs of Theorems 3.3 and 3.4 are essentially the same, and outside of obvious modifications the arguments are the same as those appearing in the proof of Lemma 3.1 of [3]. For completeness we provide here a proof of Theorem 3.4.

Proof. Let $f \in C_0^\infty(\Gamma)$. Upon multiplying each side of equation (3.11) by ψ^1/B, integrating the resulting expression over Ω and then applying the divergence theorem, we obtain after using the boundary conditions (3.12)–(3.13),

$$(\text{I}) \qquad \int_\Omega \frac{1}{B}\{(\psi^1_y)^2 + \epsilon^2 B^2 (\psi^1_x)^2\}\, dx\, dy + \alpha \int_\Gamma (\psi^1)^2 dx = \int_\Gamma f \psi^1 dx.$$

Since the first term on the left side of (I) is clearly nonnegative, it follows that

$$\alpha \int_\Gamma (\psi^1)^2 dx \le \int_\Gamma f \psi^1 dx \le \|f\|_{L^2(\Gamma)} \cdot \|\psi^1\|_{L^2(\Gamma)}.$$

It follows that

$$\|\bar{T}_1 f\|_{L^2(\Gamma)} = \|\psi^1\|_{L^2(\Gamma)} \le \frac{1}{\alpha}\|f\|,$$

and we have that $\bar{T}_1$ is bounded. That $\bar{T}_1$ is a positive operator follows also from (I) since

$$\begin{aligned}\langle \bar{T}_1 f, f\rangle_{L^2(\Gamma)} &= \int_\Gamma (\bar{T}_1 f)\cdot f dx = \int_\Gamma \psi^1 f dx\\ &\ge \alpha \int_\Gamma (\psi^1)^2 dx \ge 0.\end{aligned}$$

To see that $\bar{T}_1$ is symmetric, let $\psi^2(x,y;g)$ be the solution of (3.11)–(3.14) satisfying the boundary condition (3.13) with g replacing f. We have

$$\begin{aligned}\langle \bar{T}_1 f, g\rangle_{L^2(\Gamma)} &= \int_\Gamma \psi^1 g dx\\ &= \alpha\int_\Gamma \psi^1\psi^2 dx + \int_\Omega \frac{1}{B}\{\psi^1_y\psi^2_y + \epsilon^2 B^2 \psi^1_x\psi^2_x\} dxdy\\ &= \int_\Gamma f\psi^2 dx = \langle f, \bar{T}_1 g\rangle_{L^2(\Gamma)}.\end{aligned}$$

Now since $\bar{T}_1$ is a bounded, positive, symmetric operator defined on a dense subset of $L^2(\Gamma)$, it has a self-adjoint extension T_1 to $L^2(\Gamma)$ with the same bound. Also, T_1 is a positive operator.

To show that T_1 has no nontrivial null-space, assume that $T_1 f = 0 = \psi^1(x,0;f)$ on Γ where we assume $f \in C_0^\infty(\Gamma)$. Then

$$\begin{aligned}0 &= \langle f, T_1 f\rangle_{L^2(\Gamma)} = \langle f, \psi^1\rangle_{L^2(\Gamma)}\\ &= \alpha\langle \psi^1, \psi^1\rangle_{L^2(\Gamma)} + \int_\Omega \frac{1}{B}\{(\psi^1_y)^2 + \epsilon^2 B^2 (\psi^1_x)^2\}\, dx\, dy\\ &\ge \frac{1}{b_2}\int_\Omega \left\{\frac{1}{2}[(\psi^1)^2 + (\psi^1_y)^2] + \epsilon^2 B^2(\psi^1_x)^2\right\} dx\, dy.\end{aligned}$$

Thus, $\psi^1 = \psi^1_x = \psi^1_y = 0$ almost everywhere in Ω. That $f = 0$ follows from the equality

$$\begin{aligned}0 &= \int_\Omega \frac{1}{B}\{\psi^1_y\phi_y + B^2\epsilon^2\psi^1_x\phi_x\}\, dx\, dy + \alpha\langle \psi^1, \phi\rangle_{L^2(\Gamma)}\\ &= \int_\Gamma f\phi dx \quad \text{for all } \phi \in H^1(\Omega). \qquad \square\end{aligned}$$

Consequences of the Theorems 3.3 and 3.4 are that T and T_1 have inverses T^{-1} and T_1^{-1}, respectively, and they are also positive self-adjoint operators. We observe also that conditions (3.8) and (3.14) guarantee that the restriction to the boundary Γ of the solutions of (3.4)–(3.8) and (3.11)–(3.14), respectively, are in $L^2(\Gamma)$.

We consider the inhomogeneous boundary value problem

$$Lu = G \quad \text{in } \Omega, \tag{3.17}$$

$$Mu = g \quad \text{on } \Gamma_b, \tag{3.18}$$

$$u = 0 \quad \text{on } \Gamma, \tag{3.19}$$

$$u \in H^2(\Omega), \tag{3.20}$$

and prove the following result.

THEOREM 3.5. *Let $G, G_x \in L^2(\Omega)$ and $g \in H^1(\Gamma_b)$, and let u satisfy* (3.17)–(3.20). *Then there exist positive numbers ϵ_0 and C such that*

$$\|u_y\|^2_{L^2(\Gamma)} \leq C\{\|G\|^2_{L^2(\Omega)} + \|G_x\|^2_{L^2(\Omega)} + \|g\|^2_{H^1(\Gamma_b)}\} \tag{3.21}$$

whenever $\epsilon < \epsilon_0$.

There are two cases: (i) B is a constant function, and (ii) B is a nonconstant function. The proof of (i) is straightforward and will be omitted. We give a proof of (ii) in a series of lemmas.

LEMMA 3.1. *Let $u \in H^2(\Omega)$. Then $u_y(x,0) \in L^2(\Gamma)$ and*

$$\|u_y\|^2_{L^2(\Gamma)} \leq 2(\|u_y\|^2_{L^2(\Omega)} + \|u_{yy}\|^2_{L^2(\Omega)}).$$

Proof.

$$u_y(x,0) = u_y(x,y) + \int_y^0 u_{yy}(x,y)dy$$

and

$$|u_y(x,0)| \leq |u_y(x,y)| + \int_y^0 |u_{yy}(x,y)|dy$$
$$\leq |u_y(x,y)| + \left\{\int_{-1}^0 dy\right\}^{1/2} \left\{\int_{-1}^0 |u_{yy}(x,y)|^2 dy]\right\}^{1/2}.$$

It follows that

$$|u_y(x,0)|^2 \leq 2\left(|u_y(x,y)|^2 + \int_{-1}^0 |u_{yy}(x,y)|^2 dy\right).$$

Upon integrating both sides of the last inequality over Γ and over $\{y : 0 \geq y \geq -1\}$ we obtain the inequality of Lemma 1. □

LEMMA 3.2. *Let $u(x,y)$ be the solution of* (3.17)–(3.20). *Then there exists a constant C such that*

$$\|u_y\|^2_{L^2(\Omega)} \leq C \cdot P,$$

where P is the expression in (3.21) *enclosed by braces.*

Proof. Let u satisfy (3.17)–(3.20), and let ρ and γ be, respectively, the largest and smallest of the three numbers b_1, b_2, and b_3. Now multiply each side of (3.17) by u, integrate the resulting expression over Ω, and apply the divergence theorem to obtain the inequality

$$\int_\Omega \{u_y^2 + \epsilon^2[(Bu_x - B_x y u_y)^2 - B_x^2 y u u_y + BB_x u u_x]\}dxdy$$
$$= \int_{\Gamma_b} ugdx - \int_\Omega uGdxdy.$$

Upon applying the elementary inequalities

$$\int_{\Gamma_b} u^2 dx \leq \int_\Omega u_y^2 dxdy, \qquad \int_\Omega u^2 dxdy \leq \int_\Omega u_y^2 dxdy, \tag{3.22}$$

simplifying, and then choosing ϵ appropriately, we obtain

$$\int_\Omega \{u_y^2 + \epsilon^2 u_x{}^2\}dxdy \leq D\left\{\int_\Omega |G|^2 dxdy + \int_{\Gamma_b} |g|^2 dx\right\} \tag{3.23}$$

from which follows the conclusion of the lemma immediately. □

LEMMA 3.3. *Let $u(x,y)$ be the solution of* (3.17)–(3.20). *Then there exist positive constants C_1 and δ such that for $\epsilon < \delta$ we have*

$$\|u_{yy}\|^2_{L^2(\Omega)} \leq C_1 \cdot P$$

where P is the expression enclosed by braces in (3.21).

Proof. Differentiate the equations in (3.17)–(3.20) with respect to x, and set $W = u_x$ to obtain the boundary-value problem

$$\begin{aligned} &LW + \epsilon^2[\rho_1 W_x + \rho_2 W_y + \rho_3 W] \\ &\qquad = G_x + \epsilon^2[\rho_4 u_y + \rho_5 u_{yy}] \quad \text{in } \Omega, \end{aligned} \tag{3.24}$$

$$MW = g_x + \epsilon^2[\rho_8 W + \rho_9 u_y] \quad \text{on } \Gamma_b, \tag{3.25}$$

$$W = 0 \quad \text{on } \Gamma, \tag{3.26}$$

$$W \in H^1(\Omega), \tag{3.27}$$

where the functions $\rho_1, \rho_2, \rho_3, \rho_4$, and ρ_5 are bounded on $\overline{\Omega}$, the closure of Ω, and ρ_8 and ρ_9 are bounded in $\mathbb{R}$. Furthermore, we define $\rho_6 = B_x^2 y$ and $\rho_7 = -BB_x$. (Here, we have assumed that B and its derivatives through the third order are bounded in $\mathbb{R}$.) The constant ρ, defined by

$$\rho = \max\left\{ \max_{1 \leq i \leq 7} \left[\sup_{(x,y)\in\overline{\Omega}} |\rho_i(x,y)| \right], \max_{i=8,9} \left[\sup_{(x,y)\in\Gamma_b} |\rho_i(x,y)| \right] \right\},$$

is positive since B is a nonconstant function by assumption. We proceed as in the proof of Lemma 3.2 to obtain

$$\begin{aligned} \int_\Omega &\{W_y^2 + \epsilon^2[B^2W_x^2 - 2BB_xyW_xW_y + B_x^2y^2W_y^2 \\ &\qquad - (\rho_1 + \rho_7)W_xW - (\rho_2 + \rho_6)W_yW - \rho_3W^2]\}dxdy \\ &= -\int_\Omega W\{G_x + \epsilon^2[\rho_4 u_y + \rho_5 u]\}dxdy \\ &\quad - \int_{\Gamma_b} W\{g_x + \epsilon^2[\rho_8 W + \rho_9 u_y]\}dx. \end{aligned}$$

Employing modifications of the inequalities in (3.22) and the inequality

$$\int_\Gamma u_y^2 dx \leq 2\int_\Omega (|u_y|^2 + |u_{yy}|^2)dxdy,$$

we can find ϵ_1 greater than zero and determine constants D, C_3, and λ so that

$$\begin{aligned} &\int_\Omega (W_y^2 + \epsilon^2 W_x^2 + W^2)dxdy \\ &\quad \leq D\left\{\int_\Omega |G_x|^2 dxdy + \int_{\Gamma_b} |g_x|^2 dx\right\} \\ &\qquad + D\int_\Omega \epsilon^2(2+\epsilon^2)[u_y^2 + u_{yy}^2]dxdy \\ &\quad \leq C_3 P + \lambda\epsilon^2 \int_\Omega u_{yy}^2 dxdy = P_1 \end{aligned}$$

whenever $\epsilon < \epsilon_1$. Thus for $\epsilon < \epsilon_1$ each of the quantities $\|u_x\|^2_{L^2(\Omega)}$, $\|u_{xy}\|^2_{L^2(\Omega)}$ and $\epsilon^2\|u_{xx}\|^2_{L^2(\Omega)}$ is equal to or less than P_1. From the differential equation (3.17) it is easily shown that there exists a constant γ such that

$$|u_{yy}|^2 \leq \gamma\epsilon^4\{|u_{xx}|^2 + |u_{xy}|^2 + |u_x|^2 + |u_y|^2\} + \gamma|G|^2.$$

Upon integrating the latter expression over Ω and employing the inequalities derived above, we obtain

$$\begin{aligned}\|u_{yy}\|^2_{L^2(\Omega)} &\leq \gamma\epsilon^2(1+3\epsilon^2)P_1 + \gamma\|G\|^2_{L^2(\Omega)}\\ &\leq \gamma\{1 + C_3\epsilon^2(1+3\epsilon^2)\}P + \gamma\lambda\epsilon^4(1+3\epsilon^2)\|u_{yy}\|^2_{L^2(\Omega)}\end{aligned}$$

or equivalently,

$$[1 - \gamma\lambda\epsilon^4(1+3\epsilon^2)]\|u_{yy}\|^2_{L^2(\Omega)} \leq \gamma\{1 + C_3\epsilon^2(1+3\epsilon^2)\}P.$$

For each fixed $\beta \in (0,1)$ there exists an $\epsilon_2 > 2$ such that

$$[1 - \gamma\lambda\epsilon^4(1+3\epsilon^2)] > \beta \quad \text{whenever } 0 < \epsilon < \epsilon_2.$$

Now set $\epsilon_0 = \min\{\epsilon_1, \epsilon_2\}$. For $0 < \epsilon < \epsilon_0$ it follows that

$$\|u_{yy}\|^2_{L^2(\Omega)} \leq C_2 \cdot P,$$

where $C_2 = (\gamma/\beta)[1 + 4C_3]$. This completes the proof of the lemma. □

The proof of the theorem is now immediate.

In a sequence of lemmas we collect certain properties of the solution of the boundary-value problem

$$L_1\varphi = \varphi_{yy} + \epsilon^2 B(B\varphi_x)_x = 0 \quad \text{in } \Omega, \tag{3.28}$$

$$\varphi_y = 0 \quad \text{on } \Gamma_b, \tag{3.29}$$

$$\varphi(x,0) = h(x) \quad \text{on } \Gamma, \tag{3.30}$$

$$\varphi(x,y) \in H^1(\Omega), \tag{3.31}$$

which will be of importance in the sequel.

LEMMA 3.4. *For $h \in L^2(\Gamma)$ the function $\tilde{\varphi}$, defined by*

$$\tilde{\varphi}(\lambda, y) = \frac{\cosh(\epsilon|\lambda|(y+1))}{\cosh(\epsilon|\lambda|)}\tilde{h}(\lambda), \tag{3.32}$$

is the $\mathcal{F}_B$ transform of the solution of (3.28)–(3.31).

Proof. Let $h \in C_0^\infty(\Gamma)$. Upon applying the B-transform to (3.28)–(3.30) we obtain a boundary-value problem for an ordinary differential equation in y for $\tilde{\varphi}(\lambda, y)$. Its solution is given by (3.32). It is clear that the one-parameter family of operators $\bar{S}(y)$, $-1 \leq y \leq 0$, defined by

$$\bar{S}(y)h(x) = \mathcal{F}_B^{-1}\left[\frac{\cosh(\epsilon|\lambda|(y+1))}{\cosh(\epsilon|\lambda|)}\tilde{h}(\lambda)\right] = \varphi(x,y),$$

is bounded on $C_0^\infty(\Gamma)$ with respect to the L^2 norm. Since $C_0^\infty(\Gamma)$ is dense in $L^2(\Gamma)$, $\bar{S}(y)$ can be extended by continuity to $S(y)$ on $L^2(\Gamma)$, where $S(y)$ has the same bound as $\bar{S}(y)$. □

LEMMA 3.5. *If $h \in H^k(\Gamma)$, then $\varphi_y(x,0) \in H^{k-1}(\Gamma)$.*

Proof. From (3.32) we obtain

$$\tilde{\varphi}_y(\lambda,0) = \epsilon|\lambda| \tanh(\epsilon|\lambda|)\tilde{h}(\lambda).$$

Thus we have

$$|\tilde{\varphi}_y(\lambda,0)|^2 \leq C_1\lambda^2|\tilde{h}(\lambda)|^2 \leq C_1(1+\lambda^2)|\tilde{h}(\lambda)|^2.$$

This inequality and the assumption that $h \in H^k(\Gamma)$, together with Lemma 2.4, guarantee that $\varphi_y(x,0) \in H^{k-1}(\Gamma)$. □

LEMMA 3.6. *If $h \in H^k(\Gamma)$, then $\varphi \in H^k(\Omega)$.*

Proof.

$$\begin{aligned}
\|\varphi\|^2_{H^k(\Omega)} &= \int_\Omega \sum_{|\alpha|\leq k} \left| \frac{\partial^{\alpha_1+\alpha_2}}{\partial x^{\alpha_1}\partial y^{\alpha_2}}\varphi\right|^2 dx\, dy \\
&= \int_{\mathbb{R}}\int_{-1}^{0} \sum_{|\alpha|\leq k} \left||\lambda|^{\alpha_1}\frac{\partial^{\alpha_2}}{\partial y^{\alpha_2}}\hat{\varphi}(\lambda,y)\right|^2 dy\, d\lambda \\
&\leq C_1 \int_{\mathbb{R}}\int_{-1}^{0} \sum_{|\alpha|\leq k}\sum_{j=1}^{\alpha_1} \left||\lambda|^{j}\frac{\partial^{\alpha_2}}{\partial y^{\alpha_2}}\tilde{\varphi}(\lambda,y)\right|^2 d\lambda\, dy \\
&\leq C_2 \int_{\mathbb{R}}\int_{-1}^{0} \sum_{|\alpha|\leq k}\sum_{j=0}^{k-\alpha_2} ||\lambda|^{\alpha_2+j}\epsilon^{\alpha_2}\Upsilon_{\alpha_2}(\lambda,y)\tilde{h}(\lambda)|^2 dy\, d\lambda \\
&\leq C_3 \int_{\mathbb{R}} \sum_{|\alpha|\leq k}\sum_{j=0}^{k}\binom{k}{j} ||\lambda|^{j}\tilde{h}(\lambda)|^2 d\lambda \\
&\leq C_4 \int_{\mathbb{R}} (1+|\lambda|^2)^k|\tilde{h}(\lambda)|^2 d\lambda \\
&\leq C_5\|h\|^2_{H^k(\Gamma)} < \infty,
\end{aligned}$$

and the conclusion is established. □

The function Υ_{α_2} is defined by

$$\Upsilon_{\alpha_2}(\lambda,y) = \begin{cases} \dfrac{\cosh(\epsilon|\lambda|(y+1))}{\cosh(\epsilon|\lambda|)} & \text{if } \alpha_2 \text{ is even,} \\ \dfrac{\sinh(\epsilon|\lambda|(y+1))}{\cosh(\epsilon|\lambda|)} & \text{if } \alpha_2 \text{ is odd.} \end{cases}$$

4. Derivation of the shallow water equations. In the last section it was shown that the inverse of T is a positive self-adjoint operator in $L^2(\Gamma)$. Since the restriction to Γ of the solution ψ of (3.5)–(3.8) is in the range of T, we may solve the equation

$$Tf(x,t) = \psi(x,0,t),$$

obtaining

$$f(x,t) = T^{-1}\psi(x,0,t).$$

Upon making this substitution for f the equations (3.9)–(3.10) become

$$\epsilon\psi_{tt}(x,0,t) + (T^{-1} - \alpha I)\psi(x,0,t) = 0, \qquad t > 0; \tag{4.1}$$

$$\psi(x,0,0) = F_0(x), \qquad \psi_t(x,0,0) = F_1(x), \tag{4.2}$$

a Cauchy problem for an abstract wave equation in $L^2(\Gamma)$. From (3.7) we obtain

$$\frac{1}{B}\psi_y(x,0,t) = (T^{-1} - \alpha I)\psi(x,0,t) \quad \text{on } \Gamma \times \mathbb{R}^+. \tag{4.3}$$

In an analogous way we show that

$$\frac{1}{B}\psi^1_y(x,0,t) = (T_1^{-1} - \alpha I)\psi^1(x,0,t) \quad \text{on } \Gamma \times \mathbb{R}^+. \tag{4.4}$$

Employing (3.32) and (4.4) we can give the following explicit representation for $(T_1^{-1} - \alpha I)\psi^1(x,0,t)$.

THEOREM 4.1. *Let $\psi^1(x,y,t;f)$ be the solution of* (3.11)–(3.14). *Then*

$$(T_1^{-1} - \alpha I)\psi^1(x,0,t) = \frac{1}{B}\mathcal{F}_B^{-1}\epsilon|\lambda|\tanh(\epsilon|\lambda|)\tilde{\psi}^1(\lambda,0,t). \tag{4.5}$$

This suggests that $T_1^{-1} - \alpha I$ is defined in $L^2(\Gamma)$ by the formula

$$(T_1^{-1} - \alpha I)h(x) = \frac{1}{B}\mathcal{F}_B^{-1}\{\epsilon|\lambda|\tanh(\epsilon|\lambda|)\tilde{h}(\lambda)\}$$

for all h such that the L^2-norm of the left side is bounded. It is readily seen that this is obtained when $h \in H^1(\Gamma)$.

Circumstances are not so convenient for

$$(T^{-1} - \alpha I)\psi(x,0,t);$$

we are able to provide an implicit representation only.

THEOREM 4.2. *Let $\psi(x,y,t;f)$ be the solution of* (3.5)–(3.8). *Then*

$$(T^{-1} - \alpha I)\psi(x,0,t) = \frac{1}{B}\mathcal{F}_B^{-1}\{\epsilon|\lambda|\tanh(\epsilon|\lambda|)\tilde{\psi}(\lambda,0,t)\} + G_1(x,t), \tag{4.6}$$

where

$$G_1(x,t) = T_1^{-1}(T_1 - T)f(x,t)$$

and $\mathcal{F}_B^{-1}$ is the inverse B*-transform.*

Proof. Let $f \in C_0^\infty(\Gamma)$. From $\psi_1(x,0,t) = \psi(x,0,t) + [\psi_1(x,0,t) - \psi(x,0,t)]$ and the definitions for T and T_1 we obtain

$$T_1 f(x,t) = Tf(x,t) + (T_1 - T)f(x,t).$$

It follows that

$$f(x,t) = T_1^{-1}(T_1 f(x,t)) = T_1^{-1}(Tf(x,t)) + T_1^{-1}(T_1 - T)f(x,t).$$

Thus

$$T^{-1}\psi(x,0,t) = T_1^{-1}\psi(x,0,t) + T_1^{-1}(T_1 - T)f(x,t)$$

or

$$(T^{-1} - \alpha I)\psi(x,0,t) = (T_1^{-1} - \alpha I)\psi(x,0,t) + T_1^{-1}(T_1 - T)f(x,t),$$

and the proof is complete. □

Friedman and Shinbrot [4] have shown that the domain of $(T^{-1} - \alpha I)$ is also $H^1(\Gamma)$. We now define two operators K and K_1 in $L^2(\Gamma)$ by the equations

$$\epsilon^2 Kh = (T^{-1} - \alpha I)h \quad \text{for } h \in H^1(\Gamma),$$
$$\epsilon^2 K_1 h = (T_1^{-1} - \alpha I)h \quad \text{for } h \in H^1(\Gamma),$$

respectively. Equations (4.1)–(4.2) then become

$$w_{tt}(x,t) + \epsilon K w(x,t) = 0, \qquad t > 0; \tag{4.7}$$
$$w(x,0) = F_0(x), \qquad w_t(x,0) = F_1(x), \tag{4.8}$$

where $w(x,t)$ is the restriction to Γ, $\psi(x,0,t)$, of the solution ψ of (3.5)–(3.8).

The solution of this initial-value problem is given explicitly by the formula

$$w(x,t) = \int_0^\infty \left\{ \cos(\sqrt{\epsilon\mu}t) dG_\mu F_0(x) + \frac{\sin(\sqrt{\epsilon\mu}t)}{\sqrt{\epsilon\mu}} dG_\mu F_1(x) \right\}, \tag{4.9}$$

where the family of projections G_μ, $-\infty < \mu < \infty$, is a resolution of the identity with respect to the operator K, and where F_0 and F_1 are in the domains of K and $K^{1/2}$, or equivalently in $H^1(\Gamma)$ and $H^{1/2}(\Gamma)$, respectively.

Now employ (4.7) and the representation given in Theorem 4.1 to get

$$\begin{aligned} w_{tt} &+ \frac{1}{\epsilon B}\mathcal{F}_B^{-1}[(\epsilon|\lambda|)^2 \tilde{w}(\lambda,t)] \\ &= \frac{1}{\epsilon B}\mathcal{F}_B^{-1}[\epsilon|\lambda|(\epsilon|\lambda| - \tanh(\epsilon|\lambda|))\tilde{w}(\lambda,t)] - G_1(x,t) \\ &= G_2(x,t,w), \end{aligned} \tag{4.10}$$

or equivalently,

$$w_{tt} - \epsilon(Bw_x)_x = G_2(x,t,w) \tag{4.11}$$

since the second term on the left side of (4.10) is $-\epsilon(Bw_x)_x$. If the right side of (4.11) is small, then solutions of (4.11) will be close to solutions of

$$w_{tt}^0 - \epsilon(Bw_x^0)_x = 0.$$

The latter equations are the usual equations of the linear "shallow water" theory for time-dependent, two-dimensional flows and are generally used in the applications rather than those in (1.1)–(1.4) or (4.7). In the next section it is shown that this derivation can be put on a rigorous basis.

We define an operator K_0 on $H^2(\Gamma)$ by

$$K_0 h(x) = \frac{1}{\epsilon^2 B}\mathcal{F}_B^{-1}[(\epsilon|\lambda|)^2 \tilde{h}(\lambda)] = -(Bh_x)_x, \tag{4.12}$$

and consider the abstract Cauchy problem in $L^2(\Gamma)$,

$$w_{tt}^0 + \epsilon K_0 w^0(x,t) = 0, \qquad t \in [0,S], \tag{4.13}$$

$$w^0(x,0) = F_0(x), \qquad w_t^0(x,0) = F_1(x), \tag{4.14}$$

where S is any positive number.

From the form of the solution of this problem, analogous to the formula in (4.9), we obtain the following estimates.

LEMMA 4.1. *Let $w^0(x,t)$ satisfy the initial value problem* (4.13) *and* (4.14). *Then*

$$\|D_t^k w^0(x,t)\|_{L^2(\Gamma)} = O(\epsilon^{(k/2)-(1/2)}), \qquad k = 0,1,2,\dots; \tag{4.15a}$$

$$\|(\epsilon K_0)^k w^0(x,t)\|_{L^2(\Gamma)} = O(\epsilon^{k-(1/2)}), \qquad k = 0,1,2,\dots. \tag{4.15b}$$

Here the constant in the "O" symbol depends upon S, $\|F_0\|_{H^2(\Gamma)}$ and $\|F_1\|_{H^2(\Gamma)}$ for $0 \le k \le 2$; and depends upon S, $\|F_0\|_{H^k(\Gamma)}$ and $\|F_1\|_{H^{k-1}(\Gamma)}$ for $k > 2$.

5. Establishment of error estimates. In this section we show that the derivation given above of the "shallow water" equations can be made rigorous by obtaining an estimate for the error

$$w_t(x,t) - w_t^0(x,t)$$

that results when (4.7)–(4.8) is replaced by the initial-value problem in (4.13)–(4.14). This satisfies Friedrichs' requirement, and therefore provides a mathematical justification for the "shallow water" theory.

THEOREM 5.1.

(i) *Let w and w^0 be solutions, respectively, of the initial-value problems* (4.7)–(4.8) *and* (4.13)–(4.14);

(ii) *Let F_0 be an element of $H^2(\Gamma)$; and*

(iii) *Let F_1 be an element of $H^1(\Gamma)$.*

Then for each $S > 0$ there exists a constant C (depending upon B_3, S, $\|F_0\|_{H^2(\Gamma)}$, and $\|F_1\|_{H^1(\Gamma)}$) such that

$$\|w_t(x,t) - w_t^0(x,t)\|_{L^2(\Gamma)} \le C\sqrt{\epsilon} \quad \textit{for } t \in [0,S]. \tag{5.1}$$

Remark. The estimate in the theorem is sharp as can be seen from a consideration of (1.1)–(1.4) for the special case when the bottom is horizontal ($B(x)$ a constant function). From its explicit solution one verifies directly that the estimate in (5.1) is satisfied.

Proof. The function

$$Z(x,t) = w(x,t) - w^0(x,t)$$

satisfies the initial-value problem

$$Z_{tt}(x,t) + \epsilon K Z(x,t) = -\epsilon(K - K_0)w^0(x,t), \tag{5.2}$$

$$Z(x,0) = Z_t(x,0) = 0. \tag{5.3}$$

An application of Lemma 2.5 to the initial-value problem (5.2) and (5.3) yields the inequality

$$\begin{aligned}\|Z_t(x,t)\|_{L^2(\Gamma)} &\le \int_0^S \{\|\epsilon(K-K_0)w^0(x,t)\|_{L^2(\Gamma)}\}dt \\ &\le \int_0^S \{\|\epsilon(K_1-K_0)w^0(x,t)\|_{L^2(\Gamma)} \\ &\qquad + \|\epsilon(K-K_1)w^0(x,t)\|_{L^2(\Gamma)}\}dt \\ &= I_1 + I_2.\end{aligned}$$

We first consider I_1. Employ Lemma 2.2 to obtain

$$\begin{array}{lrl}
\text{(i)} & b_1\|\epsilon(K_1 - K_0)w^0(x,t)\|_{L^2(\Gamma)} & \le \|\epsilon B(K_1 - K_0)w^0(x,t)\|_{L^2(\Gamma)} \\
\text{(ii)} & & = \|\mathcal{F}_B^{-1}[(|\lambda|\tanh(\epsilon|\lambda|) - |\epsilon|\lambda|^2)\widetilde{w^0}(\lambda,t)]\|_{L^2(\Gamma)} \\
\text{(iii)} & & \le C_1\||\lambda|(\tanh(\epsilon|\lambda|) - \epsilon|\lambda|)\widetilde{w^0}\|_{L^2(\Gamma)} \\
\text{(iv)} & & \le C_2\|\epsilon|\lambda|^2\widetilde{w^0}\|_{L^2(\Gamma)} \\
\text{(v)} & & = C_2\|\frac{1}{\epsilon}(\epsilon|\lambda|)^2\widetilde{w^0}\|_{L^2(\Gamma)} \\
\text{(vi)} & & \le C_3\left\|\frac{1}{\epsilon B}\mathcal{F}_B^{-1}[\epsilon^2|\lambda|^2\widetilde{w^0}]\right\|_{L^2(\Gamma)} \\
\text{(vii)} & & = C_3\left\|\frac{1}{\epsilon B}\mathcal{F}_B^{-1}[\widetilde{\epsilon B K_0 w^0}]\right\|_{L^2(\Gamma)} \\
\text{(viii)} & & \le C_4\|\epsilon K_0 w^0\|_{L^2(\Gamma)} \\
\text{(ix)} & & = O(\sqrt{\epsilon}).
\end{array}$$

From this last inequality follows the existence of C_5 such that $I_1 \le C_5\sqrt{\epsilon}$. In the above analysis (iv) is obtained from (iii) by employing the inequality $\tanh x \le x$, valid for $x \ge 0$; (vi) follows from (v) by applying Lemma 2.2; (vii) is obtained from (vi) by using the definition of K_0; and (ix) follows from (viii) by applying (4.15b). We turn now to I_2. From (4.3) and (4.4) we have

$$\epsilon(K - K_1)w^0(x,t) = \frac{1}{\epsilon B}[\psi_y(x,0,t) - \psi_y^1(x,0,t)],$$

where $\psi(x,y,t)$ and $\psi^1(x,y,t)$ are the solutions of (3.5)–(3.8) and (3.11)–(3.14), respectively, which assume the value $w^0(x,t)$ on Γ. The function u, defined by

$$u(x,y,t) = \psi(x,y,t) - \psi^1(x,y,t),$$

satisfies the boundary-value problem

$$\begin{array}{lc}
(5.4) & Lu = -(L - L_1)\psi^1 = -\epsilon^2 L_2\psi^1 \quad \text{in } \Omega, \\
(5.5) & Mu = -(M - M_1)\psi^1 = -\epsilon^2 M_2\psi^1 \quad \text{on } \Gamma_b, \\
(5.6) & u = 0 \quad \text{on } \Gamma.
\end{array}$$

Bounds for ψ^1 and its derivatives play a key role in establishing an estimate for $\|u_y\|_{L^2(\Gamma)}$ where u satisfies (5.4)–(5.6). Calculation of these bounds are facilitated by the explicit formula for ψ^1, namely,

$$\psi^1(x,y,t) = \mathcal{F}_B^{-1}\left\{\frac{\cosh(\epsilon|\lambda|(y+1))}{\cosh(\epsilon|\lambda|)}\widetilde{w^0}(\lambda,t)\right\}.$$

LEMMA 5.1. *The following estimates are valid for ψ^1 and its derivatives.*

$$\begin{array}{lc}
\text{(i)} & \|D^{\alpha_1+\alpha_2}\psi^1(x,y,t)\|_{L^2(\Omega)}^2 = O(\epsilon^{2\alpha_2-1}), \\
\text{(ii)} & \|D^{\alpha_1}\psi^1(x,-1,t)\|_{L^2(\Gamma_b)}^2 = O(\epsilon^{-1}).
\end{array}$$

Proof of (i). It is easily seen that

$$\begin{aligned} D^{\alpha_2}\psi^1(x,y,t) &= \epsilon^{\alpha_2}\psi_1^1(x,y,t), \quad \text{where} \\ \psi_1^1(x,y,t) &= F_B^{-1}[|\lambda|^{\alpha_2}\Upsilon_{\alpha_2}(\lambda,y,\epsilon)\widetilde{w^0}(\lambda,t)]. \end{aligned}$$

Thus

$$\begin{aligned}
&\int_\Omega |D^{\alpha_1+\alpha_2}\psi^1(x,y,t)|^2 dx\, dy \\
&= \epsilon^{2\alpha_2}\int_{-1}^{0}\left\{\int_\Gamma |D^{\alpha_1}\psi_1^1(x,y,t)|^2 dx\right\} dy \\
&= \epsilon^{2\alpha_2}\int_{-1}^{0}\left\{\int_\Gamma \left||\lambda|^{\alpha_1}\widehat{\psi_1^1}(\lambda,y,t)\right|^2 d\lambda\right\} dy \\
&\le M_1\epsilon^{2\alpha_2}\int_{-1}^{0}\left\{\sum_{k\le\alpha_1}\int_\Gamma \left||\lambda|^{k}\widetilde{\psi_1^1}\right|^2 d\lambda\right\} dy \\
&\qquad \text{(by Lemma 2.3)} \\
&\le M_1\epsilon^{2\alpha_2}\int_{-1}^{0}\left\{\sum_{k\le\alpha_1}\int_\Gamma \left||\lambda|^{k+\alpha_2}\Upsilon_{\alpha_2}(\lambda,y,\epsilon)\widetilde{w^0}(\lambda,t)\right|^2 d\lambda\right\} dy \\
&\qquad \text{(by definition of } \psi_1^1) \\
&\le M_2\epsilon^{2\alpha_2}\int_{-1}^{0}\left\{\sum_{k\le\alpha_1}\int_\Gamma \left||\lambda|^{k+\alpha_2}\widetilde{w^0}(\lambda,t)\right|^2 d\lambda\right\} dy \\
&\qquad \text{(since } |\Upsilon_{\alpha_2}|\le 1) \\
&\le M_3\epsilon^{2\alpha_2}\sum_{k\le\alpha_1}\int_\Gamma \left||\lambda|^{k+\alpha_2}\widetilde{w^0}(\lambda,t)\right|^2 d\lambda \\
&\le M_4\epsilon^{2\alpha_2}\sum_{k\le\alpha_1}\int_\Gamma \left\|\left[\left(B\frac{\partial}{\partial x}\right)^{k+\alpha_2} w^0(x,t)\right]\right\|^2 dx \\
&\qquad \text{(by Lemmas 2.1 and 2.2)} \\
&= M_4\epsilon^{2\alpha_2}\sum_{k\le\alpha_1}\int_\Gamma \left|[(BK_0)^{(k+\alpha_2)/2}w^0(x,t)]\right|^2 dx \\
&\qquad \text{(by definition of } K_0) \\
&\le M_5\epsilon^{2\alpha_2}\sum_{k\le\alpha_1}\sum_{j\le[(k+\alpha_2)/2]}\int_\Gamma |K_0^j w^0(x,t)|^2 dx \\
&\le M_6\epsilon^{2\alpha_2}\sum_{k\le\alpha_1}\sum_{j\le[(k+\alpha_2)/2]}\epsilon^{-2j}\epsilon^{2j-1} \\
&\qquad \text{(by 4.15b)} \\
&\le M_7\epsilon^{2\alpha_2-1},
\end{aligned}$$

where the M_i's are constants and $[(k+\alpha_2)/2]$ is the greatest integer in $(k+\alpha_2)/2$. More stringent conditions must be placed on F_0 and F_1 when $|\alpha| = \alpha_1+\alpha_2 \ge 2$.

$F_0 \in D(K_0^{|\alpha|/2})$ and $F_1 \in D(K_0^{(|\alpha|-1)/2})$ are sufficient. The second part of the lemma can be established by following a similar procedure. Now from Theorem 3.5 we obtain

$$\begin{aligned}\|u_y(x,0,t)\|^2_{L^2(\Gamma)} &\le \epsilon^4[\|L_2\psi^1\|^2_{L^2(\Omega)} + \|(L_2\psi^1)_x\|^2_{L^2(\Omega)} \\ &\qquad + \|M_2\psi^1\|^2_{L^2(\Gamma_b)} + \|(M_2\psi^1)_x\|^2_{L^2(\Gamma_b)}] \\ &\le A_1\epsilon^4\left\{\int_\Omega [|\psi^1_{xxy}|^2 + |\psi^1_{xyy}|^2 \right. \\ &\qquad + |\psi^1_{xx}|^2 + |\psi^1_y|^2 + |\psi^1_x|^2]dxdy \\ &\qquad \left. + \int_{\Gamma_b} [|\psi^1_{xx}|^2 + |\psi^1_x|^2]dx\right\} \\ &\le A_2\epsilon^3,\end{aligned}$$

and from (4.3) and (4.4),

$$\begin{aligned}|\epsilon(K-K_1)w^0(x,t)| &\le \frac{1}{\epsilon b_1}|\psi_y(x,0,t) - \psi^1_y(x,0,t)| \\ &\le \frac{1}{\epsilon b_1}|u_y(x,0,t)|.\end{aligned}$$

Thus,

$$\|\epsilon(K-K_1)w^0(x,t)\|_{L^2(\Gamma)} \le \frac{C}{\epsilon}\|u_y(x,0,t)\|_{L^2(\Gamma)} \le A_3\sqrt{\epsilon},$$

and there exists a constant C_2 such that $I_2 \le C_2\sqrt{\epsilon}$. This completes the proof of Theorem (5.1). □

The proof of Theorem (1.1) follows immediately since

$$\eta(X,t) - \eta^0(X,t) = W^0_t(X,t) - W_t(X,t) = w^0_t(x,t) - w_t(x,t) = -Z_t(x,t).$$

The L^2 estimate given for the error in approximating the surface wave $\eta(X,t)$ with $\eta^0(X,t)$ is generally not good enough in the theory of water waves. A pointwise estimate is needed. Such an estimate can be obtained if the initial functions $F_0(X)$ and $F_1(X)$ and the function $B(X)$ are sufficiently smooth. The following result provides such an estimate.

THEOREM 5.2. *Let S be any positive number, let $F_0 \in H^3(\Gamma)$ and $F_1 \in H^2(\Gamma)$, and let B have bounded derivatives through the fourth order. Then there exits a constant C, depending upon B_4, S, $\|F_0\|_{H^3(\Gamma)}$, and $\|F_1\|_{H^2(\Gamma)}$ such that*

$$|\eta(X,t) - \eta^0(X,t)| \le C\epsilon^{1/2} \tag{5.7}$$

for $t \in [0,S]$.

Proof (*sketch*). Under the conditions of the theorem we can establish the inequality

$$\|w_t(x,t) - w^0_t(x,t)\|_{H^1(\Gamma)} \le C\sqrt{\epsilon}. \tag{5.8}$$

Because of the inequality in Theorem 5.1 we note that to establish the inequality in (5.8) it is sufficient to show that

$$\|w_{xt}(x,t) - w^0_{xt}(x,t)\|_{L^2(\Gamma)} \le C\sqrt{\epsilon} \tag{5.9}$$

for $t \in [0, S]$, where C depends upon S, B_4, $\|F_0\|_{H^3(\Gamma)}$, and $\|F_1\|_{H^2(\Gamma)}$. The demonstration of the inequality in (5.9) follows that employed in the proof of Theorem 5.1 with modifications caused by the fact that the occurring initial-value problems are inhomogeneous. Now setting $E(x,t) = \eta^0(x,t) - \eta(x,t) = w_t(x,t) - w_t^0(x,t)$, we obtain

$$\begin{aligned}|E(x,t)| &\le \int_\Gamma |\hat{E}(\lambda,t)|d\lambda \\ &= \int_\Gamma (1+\lambda^2)^{-1/2}(1+\lambda^2)^{1/2}|\hat{E}(\lambda,t)|d\lambda \\ &\le \left(\int_\Gamma (1+\lambda^2)^{-1}d\lambda\right)^{1/2}\left(\int_\Gamma (1+\lambda^2)|\hat{E}(\lambda,t)|^2 d\lambda\right)^{1/2} \\ &\le C_1\|E\|_{H^1(\Gamma)} \le C\epsilon^{1/2}.\end{aligned}$$

6. Concluding remarks. We return again to the operator K in (1.5) and observe that this operator returns the normal derivative $\Phi_Y(X,0,t)$ on Γ of the harmonic function Φ when it operates on $\Phi(X,0,t)$, the boundary value of Φ on Γ. Therefore, K is a Dirichlet–Neumann operator. Yosihara [16] and Craig [1], [2], while investigating the nonlinear water wave problem, described Dirichlet–Neumann operators and established many of their properties. Although in their investigations the fluids had finite depth, the bottoms were horizontal or deviated only slightly from the horizontal. For the problem investigated here, for the case B a constant (the bottom is horizontal) we can give an explicit representation for K, namely,

$$Kh(x) = \mathcal{F}^{-1}(|\lambda|\tanh(\epsilon|\lambda|)\hat{h}(\lambda)) \tag{6.1}$$

for $h \in H^1(\Gamma)$. Here $\mathcal{F}^{-1}(f)$ denotes the inverse Fourier transformation of f. For this special case the representation in (6.1) agrees with the one given by Craig [2].

Largely due to the central role played by the generalized Fourier transform $\mathcal{F}_B$, and the requirement that the range of B be a finite closed interval contained in the positive y-axis in order to establish certain important estimates, the initial-boundary value problem (1.1)–(1.4) when the bottom Γ_B of the region Ω intersects its surface (the presence of beaches) does not lend itself to treatment using our approach. We believe, however, our approach can be modified and extended to treat time-dependent three-dimensional linear flows in regions without beaches.

Acknowledgments. We acknowledge here the contributions of Marvin Shinbrot, Late Professor of Mathematics at the University of Victoria, to the early stages of development of this work. It was he who pointed out that the existing justifications of the "shallow water" theory for time-dependent motion were deficient, and suggested that this theory be given a rigorous mathematical justification.

REFERENCES

[1] WALTER CRAIG, *An existence theory for water waves, and the Boussinesq and the Korteweg deVries scaling limits*, Comm. Partial Differential Equations, 10 (1985), pp. 787–1003.

[2] ———, *Water waves, Hamiltonian systems and Cauchy integrals*, in Microlocal Analysis and Nonlinear Waves, Michael Beals, Richard B. Melrose and Jeffrey Rauch, eds., The IMA Volumes in Mathematics and its Applications, Vol. 30, Springer-Verlag, New York, 1988, pp. 37–45.

[3] AVNER FRIEDMAN AND MARVIN SHINBROT, *The initial value problem for the linearized equations of water waves*, J. Math. Mech., 17 (1967), pp. 107–180.

[4] ———, *The initial value problem for the linearized equations of water waves,* II, J. Math. Mech., 18 (1969), pp. 1177–1193.

[5] K. O. FRIEDRICHS, *On the derivation of the shallow water theory*, Appendix, in The Formation of Breakers and Bores. The Theory of Nonlinear Wave Propagation in Shallow Water and Open Channels, J. J. Stoker, Comm. Pure Appl. Math., 1 (1948), pp. 1–87.

[6] R. M. GARIPOV, *On the linear theory of gravity waves: the theory of existence and uniqueness*, Arch. Rational Mech. Anal., 24 (1967), pp. 352–362.

[7] J. KAMPE DEFERIET AND J. KOTIK, *Surface waves of finite energy*, Rational Mech. Anal., 2 (1953), pp. 577–585.

[8] TADAYOSHI KANO AND TAKAAHI NISHIDA, *Sur les Ondes de Surface de l'Eau avec une Justification Mathématique des Équations des Ondes en Eau Profonde*, J. Math. Kyoto Univ., 19 (1979), pp. 335–370.

[9] SIR HORACE LAMB, *Hydrodynamics*, Cambridge University Press, Cambridge, 1932.

[10] V. I. NALIMOV, *The Cauchy–Poisson Problem*, Dinamika Sploshn. Sredy, 18 (1974), pp. 104–210. (In Russian.)

[11] L. V. OVSJANNIKOV, *To the shallow water theory foundation*, Arch. Mech., 26 (1974), pp. 407–422.

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

[13] MARVIN SHINBROT, *Waves in shallow water*, Arch. Rational Mech. Anal., 9 (1962), pp. 234–244.

[14] S. L. SOBOLEV, *Some Applications of Functional Analysis in Mathematical Physics*, American Mathematical Society, Providence, RI, 1963.

[15] J.J. STOKER, *Water Waves*, Interscience, New York, 1957.

[16] HIDEAKI YOSIHARA, *Gravity waves on the free surface of an incompressible perfect fluid of finite depth*, Publ. Res. Inst. Math. Sci., (Kyoto Univ.), 18 (1982), pp. 49–96.

SIAM J. MATH. ANAL.
Vol. 24, No. 4, pp. 911–942, July 1993

005

NONLINEAR STABILITY AND ASYMPTOTIC BEHAVIOR OF SHEARING MOTIONS OF A NON-NEWTONIAN FLUID*

JOHN A. NOHEL† AND ROBERT L. PEGO‡

Dedicated to James B. Serrin on the occasion of his 65th birthday.

Abstract. The goal is to establish the nonlinear stability of discontinuous steady states, and study the asymptotic behavior of solutions, for the initial-boundary value problem in one space dimension governing incompressible, isothermal shear flow of a non-Newtonian fluid driven by a constant pressure gradient. The fluid is assumed to be highly elastic and viscous; the non-Newtonian contribution to the shear stress satisfies a differential constitutive law characterized by a nonmonotone relation between the total *steady* shear stress and shear strain-rate that results in steady states having, in general, discontinuities in the strain rate. In a regime where Reynolds number is small compared to Deborah number, it is shown that every solution tends to a steady state as $t \to \infty$, and steady states that are nonlinearly stable, in a precise sense, are identified.

Key words. spurt, discontinuous steady states, delta-approximate solutions, gradient-like flow, Morse decomposition

AMS subject classifications. 34D20, 35B35, 35B45, 35B65, 35F25, 73F15, 76A10

1. Introduction.

In this paper we study issues of stability and asymptotic behavior for pressure-driven shear flows of an incompressible non-Newtonian fluid under isothermal conditions. We consider flow in the direction of the y-axis in the channel $-1 \le x \le 1$, and symmetric about the centerline $x = 0$; thus the flow variables are functions of x and time t. In dimensionless units, the dynamic equations governing the flow are

$$\text{(JSO)} \qquad \begin{aligned} \alpha v_t &= \left(\varepsilon v_x + \sigma + \bar{f}x\right)_x, \\ \sigma_t &= -\sigma + (Z+1)\,v_x, \\ Z_t &= -Z - \sigma v_x, \end{aligned}$$

for $-1 \le x \le 0$, $t > 0$, subject to the boundary conditions for $t > 0$:

$$\text{(BC)} \qquad v(-1,t) = 0, \qquad v_x(0,t) = 0, \qquad \sigma(0,t) = 0.$$

Here v represents fluid velocity, σ the polymer contribution to the shear stress, and Z is proportional to the first normal stress difference. The (constant) pressure gradient driving the flow is $\bar{f} > 0$, ε is the ratio of Newtonian viscosity to shear viscosity (scaled by relaxation time), and α is the ratio of Reynolds number to Deborah number. The second and third equations in (JSO) arise from the Johnson–Segalman–Oldroyd differential constitutive equations with a single relaxation time, that are assumed to

*Received by the editors February 27, 1992; accepted for publication (in revised form) November 18, 1992. This work was supported by National Science Foundation grants DMS 8907264 and DMS 8902422.

†Department of Mathematics and Center for the Mathematical Sciences, University of Wisconsin, Madison, Wisconsin 53705 and Forschungsinstitut für Mathematik, ETH-Zürich, CH-8092, Zürich, Switzerland.

‡Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742.

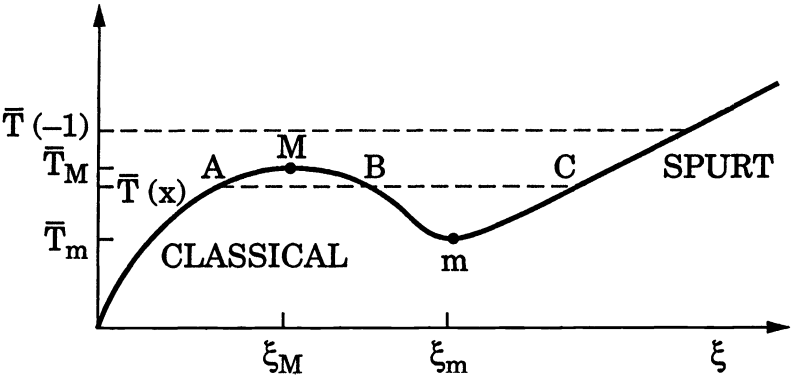


FIG. 1. ω *vs.* ξ.

govern the class of fluids under study. We refer to [8] for a derivation of the system (JSO) starting from balance laws and constitutive relations in three dimensions, and for references to relevant literature.

Throughout this paper, it is convenient to study the system (JSO) in the equivalent form:

$$\text{(JSO}_1\text{)}\qquad \begin{aligned} \alpha S_t &= \varepsilon S_{xx} + \alpha\sigma_t, \\ \sigma_t &= -\sigma + \frac{(Z+1)\left(S+\overline{T}-\sigma\right)}{\varepsilon}, \\ Z_t &= -Z - \frac{\sigma\left(S+\overline{T}-\sigma\right)}{\varepsilon}, \end{aligned}$$

where

$$\text{(1.1)}\qquad S := \varepsilon v_x + \sigma + \overline{f}\,x, \qquad \overline{T} := -\overline{f}\,x.$$

The boundary conditions become

$$\text{(BC}_1\text{)}\qquad S_x(-1,t) = 0, \qquad S(0,t) = 0, \qquad \sigma(0,t) = 0,$$

and the initial conditions are

$$\text{(IC}_1\text{)}\qquad S(x,0) = S_0(x), \qquad \sigma(x,0) = \sigma_0(x), \qquad Z(x,0) = Z_0(x).$$

The steady states $(\overline{S}(x), \overline{\sigma}(x), \overline{Z}(x))$ of (JSO$_1$) satisfy the following relations (see [8]):

$$\text{(1.2)}\qquad \overline{\sigma}(x) = \frac{\overline{v}_x}{1+\overline{v}_x^2}, \qquad \overline{Z}(x) + 1 = \frac{1}{1+\overline{v}_x^2},$$

$$\text{(1.3)}\qquad \overline{S}(x) = \varepsilon\overline{v}_x + \frac{\overline{v}_x}{1+\overline{v}_x^2} + \overline{f}\,x \equiv 0.$$

Thus the steady strain rate $\overline{v}_x(x)$ satisfies the equation $\omega(\overline{v}_x) = \overline{T} = -\overline{f}\,x$, where $\overline{T}$ is the total steady shear stress, and where

$$\text{(1.4)}\qquad \omega(\xi) = \frac{\xi}{1+\xi^2} + \varepsilon\xi, \qquad -\infty < \xi < \infty.$$

For $\varepsilon < \frac{1}{8}$, the function ω is not monotone (see Fig. 1). In this case, if $\overline{f}$ is sufficiently large, there are multiple steady states $\overline{v}_x(x)$ satisfying (1.3), which are discontinuous in x. The resulting steady velocity profiles $\overline{v}(x)$ have kinks at points where $\overline{v}_x$ is discontinuous. (For an example with one kink, see Fig. 2.)

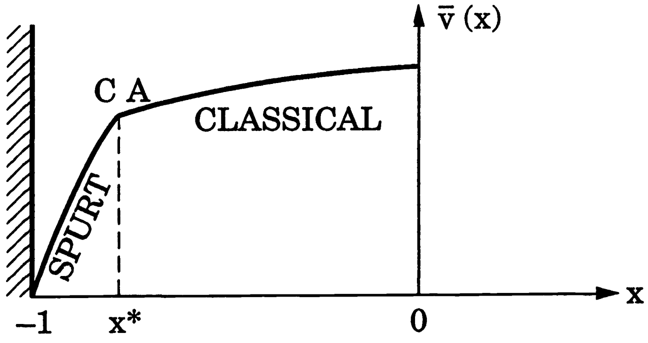


FIG. 2. *Velocity profile with a kink;* $\omega(\overline{v}_x(x)) = \overline{T}$.

Such steady states play a key role in one possible explanation of experimentally observed spurt and related phenomena that was proposed by Malkus, Nohel, and Plohr in [8]. The experiments involve highly elastic and viscous fluids, for which α and ε are both small, with $\varepsilon \sim 10^{-2}$ or 10^{-3}, and α is 7 to 10 orders of magnitude smaller than ε. The explanation in [8] is based on analytical results for the approximation of the governing system (JSO$_1$) when $\alpha = 0$; these results are summarized in Propositions 2.3 and 2.4 below. The results in [8] suggest that the spurt mechanism is a bulk material property of the viscoelastic fluid, rather than an an adhesive property as currently believed by many rheologists. As a consequence of the nonmonotone shear stress/strain-rate relation in Fig. 1, the polymeric fluid changes state in a thin layer near the boundary occuring whenever $\overline{T} > \overline{T}_M$, and giving the appearance of a "wall slip" layer; furthermore, this layer exhibits complicated dynamics.

The main focus of this paper is to show that in a regime where $\varepsilon < \frac{1}{8}$ is fixed and $\alpha > 0$ is sufficiently small, the dynamics of the full system (JSO$_1$) is similar to that generated by the approximate problem with $\alpha = 0$. We do so by studying the nonlinear stability of discontinuous steady states of (JSO$_1$) with respect to perturbations of initial data, and the large-time asymptotic behavior of solutions in general.

Precise statements of our main results appear in §3. But loosely speaking, we prove three results.

1. On any fixed interval $0 \le t \le T$ the solution (S, σ, Z) of (JSO$_1$) converges as $\alpha \to 0$ to the corresponding solution for $\alpha = 0$, for which $S = 0$ and the second and third equations of (JSO$_1$) reduce to a pair of quadratic, autonomous ordinary differential equations (ODEs) in which x enters only as a parameter through $\overline{T}$. (The case $\alpha = 0$ is thoroughly analyzed in [8]; a summary of those results will be given in §2.)

2. If α is small (in particular $\alpha\varepsilon^{-4}$ should be small), then any discontinuous steady state, for which $\omega(\overline{v}_x)$ takes values only on the strictly increasing portions of the graph of $\omega(\xi)$ (excluding a neighborhood of the max and min), is nonlinearly stable, in a sense made precise in §3. Stability holds when the perturbation from steady state in S is small in $H^1(-1,0)$ and the perturbations in σ, Z are small in $L^1(-1,0)$, and are bounded pointwise by some large constant. Hence, there are smooth initial data in the basin of Lyapunov stability of any such discontinuous steady state. Figure 3 illustrates the effect of such a perturbation on the strain rate $\overline{v}_x$ obtained from the kinked velocity profile in Fig. 2.

3. If α is sufficiently small, every solution of (JSO$_1$) converges as $t \to \infty$ to *some* steady state solution (possibly discontinuous, possibly unstable) with

$$S \to 0 \text{ in } H^1(-1,0), \qquad (\sigma(x,t), Z(x,t)) \to (\overline{\sigma}(x), \overline{Z}(x)) \tag{1.5}$$

for each x. We can guarantee that for some smooth solutions, the corresponding asymptotic state is discontinuous in x, and the convergence in (1.5) is not uniform; such solutions contain "transition layers." In general, we are unable to precisely identify the limiting steady solution for a given set of initial data.

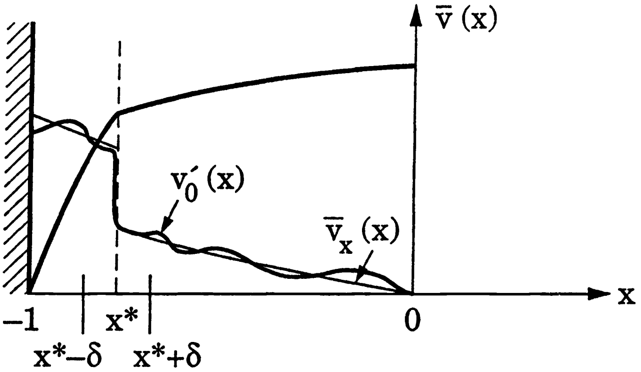


FIG. 3. *Smooth perturbation of velocity gradient.*

2. Summary of previous results.

Global existence of solutions to ($\mathrm{JSO_1}$) with the boundary conditions ($\mathrm{BC_1}$) and initial conditions ($\mathrm{IC_1}$) for arbitrary α and $\varepsilon > 0$ follows from a general result established in [9]. Somewhat different existence results have been obtained by Guillopé and Saut [4], [5]. We have the following.

THEOREM 2.1. (a) *Assume $S_0 \in H^s$, for some $s > \frac{3}{2}$, and σ_0, $Z_0 \in C^1$. Then there exists a unique* classical solution *of* ($\mathrm{JSO_1}$), ($\mathrm{BC_1}$), ($\mathrm{IC_1}$) *on* $[-1,0] \times [0,\infty)$ *having the regularity*

$$\begin{aligned} S &\in C([0,\infty), C^1) \cap C((0,\infty), H^2), \\ S_t &\in C((0,\infty), C^1), \\ \sigma,\ Z &\in C^1([0,\infty), C^1). \end{aligned}$$

(b) *Assume $S_0 \in H^1$, and σ_0, $Z_0 \in L^\infty$. Then there is a unique semiclassical solution on $[-1,0] \times [0,\infty)$ (possibly having discontinuities in the stress components) with*

$$\begin{aligned} S &\in C([0,\infty), H^1) \cap C((0,\infty), W^{2,\infty}), \\ S_t &\in C((0,\infty), H^s) \quad \textit{for all } s < 2, \\ \sigma,\ Z &\in C^1([0,\infty), L^\infty). \end{aligned}$$

Given any bounded measurable representatives σ_{0*}, Z_{0*} of the equivalence classes σ_0, $Z_0 \in L^\infty$, there exist unique bounded measurable functions $\sigma_*(x,t)$, $Z_*(x,t)$, representing $\sigma(\cdot,t)$, $Z(\cdot,t)$ for each $t > 0$, such that the map $t \mapsto \sigma_*(x,t)$, $Z_*(x,t)$ is C^1 for $t \geq 0$, and (S_*, σ_*, Z_*) satisfy the second and third equations of ($\mathrm{JSO_1}$) *for $t > 0$, where S_* is the unique continous representative of S.*

Throughout this paper, function spaces such as H^s, C^1, $W^{2,\infty}$, L^∞ refer to the interval $[-1,0]$. The class of solutions in Theorem 2.1(b) includes the discontinuous steady states mentioned in the introduction. In what follows, we will identify a solution (S,σ,Z) in case (b) with some representative (S_*,σ_*,Z_*) as described above; then the second and third equations in ($\mathrm{JSO_1}$) are satisfied pointwise for all $t > 0$. We remark that, as shown in [9], discontinuities in the stress components (σ, Z) for such solutions can neither be created nor destroyed in finite time. This is so because S is continuous, and solutions of the ODEs in (JSO$_1$) depend continuously on parameters both forward and backward in time.

Since (JSO$_1$) is endowed with the identity

$$\text{(L)} \qquad \frac{d}{dt}\left[\sigma^2 + (Z+1)^2\right] = -2\left[\sigma^2 + \left(Z + \frac{1}{2}\right)^2 - \frac{1}{4}\right],$$

it follows that σ and Z are globally bounded in (x,t), and these bounds are independent of α and ε. But although solutions of (JSO$_1$) exist globally, it has not been shown that S remains globally bounded in (x,t). Observe that identity (L) is independent of $\alpha, \varepsilon, \overline{T}$.

Nohel, Pego, and Tzavaras [9] studied the nonlinear stability of discontinuous steady states and asymptotic behavior as $t \to \infty$, for a model problem obtained from (JSO$_1$) by freezing Z at its steady state value $Z + 1 = 1/(1+v_x^2)$. The resulting model system has the form

$$\text{(M)} \qquad \begin{aligned} \alpha S_t &= \varepsilon S_{xx} + \alpha\sigma_t, \\ 0 &= \sigma_t + \sigma + g\left(\frac{(S + \overline{T} - \sigma)}{\varepsilon}\right), \end{aligned}$$

where $g(\xi) = \xi/(1+\xi^2)$. The steady states $(\overline{S}(x), \overline{\sigma}(x))$ of (M) satisfy $\overline{S}(x) = 0$, and, with $\overline{v}_x = (\overline{T} - \overline{\sigma})/\varepsilon$,

$$\varepsilon\overline{v}_x + g(\overline{v}_x) = \overline{T}.$$

System (M) admits energy estimates which imply that $S \in L^\infty([-1,0] \times [0,\infty))$ and $S(x,t) \to 0$ as $t \to \infty$, uniformly in x. The analysis of stability and asymptotic behavior reduces to studying invariant sets of the second equation in (M), a single ODE with x as parameter, forced by S with S small.

The results of [9] offered evidence that the discontinuous steady states of (JSO$_1$) are relevant to the study of the asymptotic behavior of smooth solutions. For system (M) with $\alpha = \varepsilon = 1$ (though this is not an essential assumption), it was shown that: (i) Every solution of (M) converges as $t \to \infty$ to some steady state. (ii) Discontinuous steady states satisfying $\varepsilon + g'(\overline{v}_x) \geq c_0 > 0$ for some c_0 are nonlinearly stable in the following sense: Restrict initial data (S_0, σ_0) such that the initial values are close to $(0, \overline{\sigma}(x))$ except on the union $\mathcal{U}$ of small subintervals centered around points $x_1, \ldots, x_n$ at which $\overline{v}_x$ is discontinuous. Then we have the following.

THEOREM 2.2. *Let $(0, \overline{\sigma}(x))$ be a steady state solution of* (M) *as described above, with a finite number of discontinuities, and satisfying*

$$\varepsilon + g'(\overline{v}_x(x)) \geq c_0 > 0 \quad \textit{for a.e. } x \in [-1,0]$$

for some positive constant c_0. If the measure of $\mathcal{U}$ is sufficiently small, there is a positive constant δ_0 depending on $\mathcal{U}$ such that, if $\delta < \delta_0$, then for any initial data $(S_0(x), \sigma_0(x))$ satisfying

$$\sup_{0 \leq x \leq 1} |S_0(x)| < \delta, \qquad \int_0^1 S_{0x}(x)^2\, dx < \frac{\delta^2}{2},$$

$$|\sigma_0(x) - \overline{\sigma}(x)| < \delta, \qquad x \in [0,1] \setminus \mathcal{U},$$

the corresponding solution $(S(x,t),\sigma(x,t))$ approaches the steady state $(0,\overline{\sigma}(x))$ as $t \to \infty$, in the sense that

$$S(x,t) \to 0 \quad \textit{uniformly},$$
$$\sigma(x,t) \to \overline{\sigma}(x) \quad \textit{for all } x \in [0,1] \setminus \mathcal{U}.$$

The main results of this paper, Theorems 3.4 and 3.5 below, are generalizations of Theorem 2.2 from the model problem (M) to the system (JSO$_1$).

The limiting case $\alpha = 0$ in (JSO$_1$) was studied by Malkus, Nohel, and Plohr in [8]. Their analysis of this case predicts spurt as well as latency, shape memory, and hysteresis under cyclic loading and unloading. Putting $\alpha = 0$ in (JSO$_1$), one has $S = 0$ and the second and third equations reduce to the system of quadratic ODEs parametrized by x:

$$\text{(Q)} \qquad \begin{aligned} \sigma_t &= -\sigma + \frac{(Z+1)\left(\overline{T}-\sigma\right)}{\varepsilon}, \\ Z_t &= -Z - \frac{\sigma\left(\overline{T}-\sigma\right)}{\varepsilon}. \end{aligned}$$

(Recall $\overline{T} = -\overline{f}x$.) The steady states of (Q) are identical with the steady states of (JSO$_1$) for $\alpha > 0$, but the dynamics of solutions of (Q) can be studied independently for each x in $[-1,0]$. Considering $\overline{T} = -\overline{f}x$ as a fixed parameter in (Q), we summarize the results of [8] as follows.

1. For each $\overline{T} \geq 0$, system (Q) has no periodic or homoclinic orbits, and every solution converges as $t \to \infty$ to some critical point.

2. The critical points of (Q) lie in the fourth quadrant of the (σ, Z)-phase plane, at the intersections of the circle

$$\Gamma := \left\{ (\sigma, Z) \mid \sigma^2 + \left(Z + \frac{1}{2}\right)^2 = \frac{1}{4} \right\}$$

and the parabola $Z = \sigma(\sigma - \overline{T})/\varepsilon$.

The character of the critical points is described as follows: In Fig. 1, denote the coordinates of the local maximum M and the local minimum m by $(\xi_M, \overline{T}_M)$ and $(\xi_m, \overline{T}_m)$, respectively; $\overline{T}_m$ and $\overline{T}_M$ are the critical values of the function $\omega(\xi)$, with $\overline{T}_m < \overline{T}_M$. There are three cases.

(i) If $0 < \overline{T} < \overline{T}_m$, there is a single critical point $A = (\sigma_A, Z_A)$ which is a globally attracting node.

(ii) If $\overline{T} > \overline{T}_M$, there is a single, globally attracting spiral point $C = (\sigma_C, Z_C)$.

(iii) If $\overline{T}_m < \overline{T} < \overline{T}_M$, there are three critical points A, B, C (see Fig. 4). A is an attracting node, B is a saddle point, and C is generally an attracting spiral point. (But for $\overline{T}$ close to $\overline{T}_m$, C is an attracting node.)

Two saddle-node bifurcations occur as $\overline{T}$ varies: As $\overline{T} \to \overline{T}_m$ from above, points B and C coalesce, and as $\overline{T} \to \overline{T}_M$ from below, points A and B coalesce. The set of critical points of (Q) in the full $(\sigma, Z, \overline{T})$ parameter space is a single smooth curve; this set is visualized in Fig. 5.

3. The asymptotic behavior of the solutions of (Q) is completely characterized as follows: For $\overline{T} < \overline{T}_m$ or $\overline{T} > \overline{T}_M$, every solution tends to the unique critical point, A or C, respectively. For $\overline{T}_m < \overline{T} < \overline{T}_M$, the behavior of solutions is described by Proposition 3.5 in [8], reproduced here as the following.

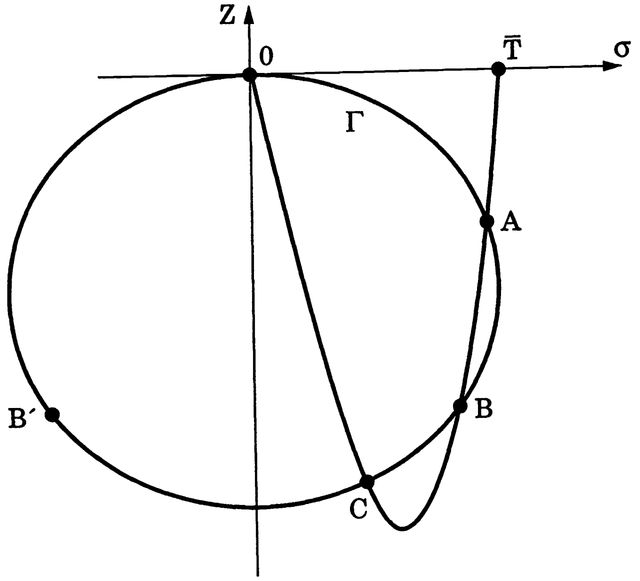


FIG. 4. *Critical points of* (Q) *in case* (iii).

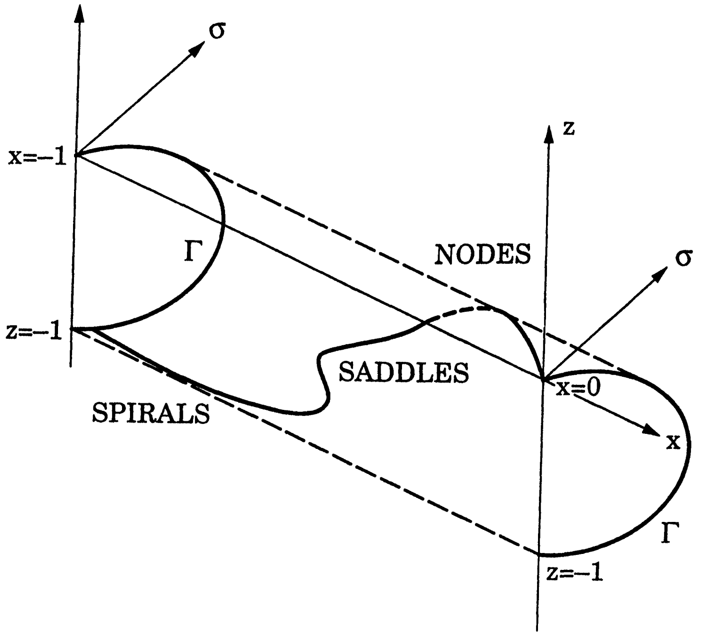


FIG. 5. *Manifold of equilibria of* (Q).

PROPOSITION 2.3. *The basin of attraction of A; i.e., the set of points that flow toward A as $t \to \infty$, comprises those points on the same side of the stable manifold of B as A; points on the other side are in the basin of attraction of C. Moreover, the arc of the circle Γ, through the origin, between $B = (\sigma_B, Z_B)$ and its reflection $B' = (-\sigma_B, Z_B)$, is contained in the basin of attraction of A.*

These results for (Q) have the following consequences for solutions $(0, \sigma, Z)$ of (JSO_1) with $\alpha = 0$, for $(x,t) \in [-1,0] \times [0,\infty)$.

PROPOSITION 2.4. *Consider* (JSO_1) *with* $\alpha = 0$.

(i) *The asymptotic behavior of any given solution may be completely characterized. For each x in $[-1,0]$, we have $(\sigma(x,t), Z(x,t)) \to A$, B or C as $t \to \infty$, according as to whether $(\sigma_0(x), Z_0(x))$ lies in the basin of attraction of A, on the stable manifold of B, or in the basin of attraction of C.*

(ii) *The stable steady states $(0, \overline{\sigma}, \overline{Z})$ of* (JSO$_1$) *with $\alpha = 0$ are those for which*

$$(\overline{\sigma}(x), \overline{Z}(x)) = A \quad \text{or} \quad C \quad \text{for a.e. } x \text{ in } [-1,0].$$

3. Statements of main results. Our main goal in this paper is to extend as far as possible the results of Proposition 2.4 for $\alpha = 0$ to the case when $\alpha > 0$ is small. Throughout this paper, ε is fixed, $0 < \varepsilon < \frac{1}{8}$. First, we study taking the limit $\alpha \to 0$ in a finite time interval $0 \le t \le T$. It will be convenient to write (JSO$_1$) in the more compact form

$$\begin{aligned} S_t &= \frac{\varepsilon}{\alpha} S_{xx} + a(x,u)S + b(x,u), \\ u_t &= G(x,u) + H(x,u)S, \end{aligned} \tag{3.1}$$

where $u = (\sigma, Z)$ and the components of $G = (G_1, G_2)$ and $H = (H_1, H_2)$ are

$$\begin{aligned} G_1(x,u) &= -\sigma + \frac{(Z+1)(\overline{T}-\sigma)}{\varepsilon}, & H_1(x,u) &= \frac{(Z+1)}{\varepsilon}, \\ G_2(x,u) &= -Z - \frac{\sigma(\overline{T}-\sigma)}{\varepsilon}, & H_2(x,u) &= -\frac{\sigma}{\varepsilon}, \end{aligned} \tag{3.2}$$

and

$$a(x,u) = H_1(x,u), \qquad b(x,u) = G_1(x,u). \tag{3.3}$$

In what follows, what is important about system (3.1) is that an a priori bound exists for $|u(x,t)|$ independent of S, namely, for some constant M_u independent of α (and ε), depending only on the initial data $u(x,0) = u_0(x)$,

$$\sup_{\substack{-1 \le x \le 0 \\ t \ge 0}} |u(x,t)| \le M_u. \tag{3.4}$$

Moreover, system (3.1) is linear in S, and the functions a, b, G, H and their Lipschitz constants with respect to u are bounded by some constant L/ε, where L is independent of α (and ε).

Our first result, proved in §4, bounds the H^1-norm of S globally in time for small α.

PROPOSITION 3.1. *Let (S,u) be a solution of (3.1) with the properties given by Theorem 2.1. Put $\lambda = \pi^2/4$ and let $\tilde{K}$ be the constant*

$$\tilde{K} = \int_0^{1/2\lambda} (2et)^{-1/2} e^{\lambda t/2}\, dt + \int_{1/2\lambda}^{\infty} \lambda^{1/2} e^{-\lambda t/2}\, dt.$$

If $\alpha \varepsilon^{-2} L \tilde{K} < \frac{1}{2}$, then

$$\|S(\cdot,t)\|_{H^1} \le 2\|S_0\|_{H^1} e^{-\varepsilon\lambda t/2\alpha} + \frac{4\alpha}{3\varepsilon} \sup_{0 \le \tau \le t} \|b(\cdot, u(\cdot,\tau))\|_{L^2}. \tag{3.5}$$

Immediate consequences of estimate (3.5) are the following.

COROLLARY 3.2. *Under the hypotheses above, there are constants* C_0, C_1, *depending only on initial data, such that*

$$\|S(\cdot,t)\|_{H^1} \le C_0 e^{-\varepsilon\lambda t/2\alpha} + C_1\alpha, \qquad t \ge 0, \tag{3.6}$$

$$\|S(\cdot,t)\|_{H^1} \le C_1\alpha, \qquad t \ge t_0(\alpha), \tag{3.7}$$

where

$$t_0(\alpha) = \frac{2\alpha\log\alpha^{-1}}{\varepsilon\lambda}, \quad C_1 = 2\|S_0\|_{H^1} + \frac{4}{3}M_u L\varepsilon^{-2}, \quad C_0 = 2\|S_0\|_{H^1}.$$

THEOREM 3.3. *Let* $(S,\sigma,Z)(x,t,\alpha)$ *denote the solution of* (JSO_1), *satisfying* (BC_1) *and* (IC_1). *Let* $(0,\overset{\circ}{\sigma}(x,t),\overset{\circ}{Z}(x,t))$ *be the solution of* (JSO_1) *with* $\alpha = 0$ *and initial data* $\overset{\circ}{\sigma}(x,0) = \sigma_0(x)$, $\overset{\circ}{Z}(x,0) = Z_0(x)$. *Fix* $T > 0$. *As* $\alpha \to 0$, *we have*

$$\begin{aligned} S(x,t,\alpha) &= 0 + O(\alpha), && t_0(\alpha) \le t \le T,\\ \sigma(x,t,\alpha) &= \overset{\circ}{\sigma}(x,t) + O(\alpha), && 0 \le t \le T,\\ Z(x,t,\alpha) &= \overset{\circ}{Z}(x,t) + O(\alpha), && 0 \le t \le T, \end{aligned} \tag{3.8}$$

uniformly with respect to x *on* $[-1,0]$.

The estimates for σ, Z are proved at the end of §4.

Our next result, proved in §5, concerns the stability of discontinuous steady states of (JSO_1) for small α. Recall that A, B, C are the critical points of (Q) in the (σ, Z)-phase plane, and depend on $\overline{T}$.

THEOREM 3.4. *Let* $h_0 > 0$. *Then there exist positive constants* α_0, δ_0, *and* K *so that the following holds for* $\alpha < \alpha_0$. *Let* $(0,\overline{u}(x))$ *be a steady state solution of* (3.1) *such that*

$$\overline{u}(x) = (\overline{\sigma}(x),\overline{Z}(x)) = \begin{cases} A & \textit{when } 0 \le \overline{T} \le \overline{T}_m + h_0,\\ A \text{ or } C & \textit{when } \overline{T}_m + h_0 \le \overline{T} \le \overline{T}_M - h_0,\\ C & \textit{when } \overline{T}_M - h_0 \le \overline{T} \end{cases}$$

for $-1 \le x \le 0$. *For* $0 < \delta < \delta_0$, *assume that*

$$\|S_0\|_{H^1} \le \delta \quad \textit{and} \quad |u_0(x) - \overline{u}(x)| \le \delta \quad \textit{for } x \in [-1,0]\setminus\mathcal{U},$$

where $\mathcal{U} \subset [-1,0]$ *is a set with* $\mathrm{meas}(\mathcal{U}) \le \delta^2$. *Then for all* $t > 0$, *the solution of* (3.1) *with these initial data satisfies*

$$\|S(\cdot,t)\|_{H^1} \le K\delta \quad \textit{and} \quad |u(x,t) - \overline{u}(x)| \le K\delta \quad \textit{for } x \in [-1,0]\setminus\mathcal{U}.$$

Remark 1. Theorem 3.4 permits the steady state stress components $\overline{u}(x)$ to exhibit an arbitrary pattern of discontinuities in the x-interval, where $\overline{u}(x) = A$ or C is possible.

Remark 2. Theorems 3.4 and 2.1 guarantee that such steady states $(0,\overline{u})$ are stable in the space $H^1 \times L^\infty$. But if $\|u_0 - \overline{u}\|_{L^\infty} < \delta$ is small, the set $\mathcal{U}$ will be empty, and u_0 will be discontinuous wherever $\overline{u}$ is. So the question arises: Can discontinuous steady states attract smooth, classical solutions?

Remark 3. The answer to this question is yes. Regardless of the pattern of discontinuities of $\overline{u}(x)$, clearly there are always smooth initial data (S_0, u_0) satisfying the hypotheses of Theorem 3.4 for arbitrary $\delta > 0$. It suffices that S_0 be small in H^1 and that $u_0 - \overline{u}$ be pointwise bounded, and small in L^1, since by Chebyshev's inequality,

$$\|u_0 - \overline{u}\|_{L^1} \leq \delta^3 \quad \text{implies meas}\{x : |u_0(x) - \overline{u}(x)| > \delta\} \leq \delta^2.$$

For u_0 smooth and $\overline{u}$ discontinuous, the "transition set" $\mathcal{U}$ will be nonempty; but the conclusion of Theorem 3.4 guarantees that $u(x,t) - \overline{u}(x)$ stays small for every fixed x for which it was small initially.

Remark 4. Using Theorem 3.5 below, it is possible to guarantee that for smooth initial data that satisfy the hypotheses of Theorem 3.4 with discontinuous $\overline{u}$, there is pointwise, nonuniform, convergence as $t \to \infty$ to an asymptotic state $(0, u_\infty(x))$ which is discontinuous, with $u_\infty(x) = \overline{u}(x)$ for $x \in [-1, 0] \setminus \mathcal{U}$.

Remark 5. When starting from rest with $(S_0, u_0) = (0, 0)$, and $\overline{f} > \overline{T}_M$, the solution of (JSO$_1$) with $\alpha = 0$ converges as $t \to \infty$ to the "top jumping" steady state $(0, \overline{u}(x))$ with

$$\overline{u}(x) = (\overline{\sigma}(x), \overline{Z}(x)) = \begin{cases} A & \text{for } x \text{ where } 0 \leq \overline{T} < \overline{T}_M, \\ C & \text{for } x \text{ where } \overline{T}_M < \overline{T} \end{cases}$$

(see [8]). Theorem 3.4 does not address the stability of this solution, nor that of corresponding "bottom jumping" solutions with $\overline{u}(x) = C$ when $\overline{T}$ is near $\overline{T}_m$; the crucial estimate (5.3) in §5 is not valid for such solutions.

Remark 6. It is seen in the proof that we require $K\alpha\varepsilon^{-4} < \frac{1}{2}$, where K is a fixed constant independent of α and ϵ.

The last issue we address for (JSO$_1$) is the asymptotic behavior of arbitrary solutions, when $\alpha > 0$ is small. Our ultimate goal is to prove the following.

THEOREM 3.5. *For any constant $M > 0$, there exists $\alpha_0 > 0$ such that if $0 < \alpha < \alpha_0$, then any solution (S, σ, Z) of* (JSO$_1$), (BC$_1$), (IC$_1$), *with initial values satisfying*

$$\|S_0\|_{H^1} + \|\sigma_0\|_{L^\infty} + \|Z_0\|_{L^\infty} \leq M,$$

converges as $t \to \infty$ to some steady state $(0, \overline{\sigma}(x), \overline{Z}(x))$, in the sense that as $t \to \infty$,

$$\begin{aligned} &\|S(\cdot, t)\|_{H^1} \to 0, \\ &\sigma(x,t) \to \overline{\sigma}(x) \quad \textit{for each } x \in [-1, 0], \\ &Z(x,t) \to \overline{Z}(x) \quad \textit{for each } x \in [-1, 0]. \end{aligned}$$

The proof of this theorem requires several steps, which we now outline. The main point of view underpinning the analysis is that for small α, the bound on S from Proposition 3.1 allows us to consider the evolution of $u = (\sigma, Z)$ as given *approximately* by the ODEs in (Q) for each $x \in [-1, 0]$ independently.

DEFINITION 3.6. For $\delta > 0$, a *δ-approximate solution* of the system (Q), written as

$$\dot{u} = G(x, u) \tag{3.9}$$

with x fixed, is a C^1-function $w : [0, \infty) \mapsto R^2$ such that

$$|\dot{w} - G(x, w)| \leq \delta, \qquad 0 \leq t < \infty,$$

where $|\cdot|$ denotes a vector norm in R^2.

Recall from §2 that in [8] it was shown that the system (Q) has no periodic of homoclinic orbits, and the critical points are always hyperbolic (except exactly when $\overline{T} = \overline{T}_m$ or $\overline{T}_M$, where the saddle-node bifurcations occur). We will see that for small δ, it follows that any δ-approximate solution of (3.9) must approximately stabilize as $t \to \infty$. One result we will use in this direction says that a δ-approximate solution w must enter a region where $|G(x,w)|$ is small, within a finite time T independent of x (although u may later leave this region).

PROPOSITION 3.7. *Let $\mathcal{E}$ denote the manifold of critical points of* (Q)*, given by*

$$\mathcal{E} = \{(x,u) \mid -1 \le x \le 0 \text{ and } G(x,u) = 0\}. \tag{3.10}$$

Let $\mathcal{N}$ be any neighborhood of $\mathcal{E}$ in $[-1,0] \times R^2$, and let $M_u > 0$. Then there exists $\delta_0 > 0$ and $T > 0$ such that if $0 \le \delta < \delta_0$, and if $u(x,t)$ satisfies (3.4) *and is such that $u(x,\cdot)$ is a δ-approximate solution of* (3.9) *for each x, then for each $x \in [-1,0]$ there exists $t < T$ with*

$$u(x,t) \in \mathcal{N}. \tag{3.11}$$

Since the proof is short and elementary, we give it here in order not to interrupt the development that follows.

Proof of Proposition 3.7. Choose $h > 0$ so that, defining

$$\mathcal{B}_h(\mathcal{E}) := \{(y,w) \in [-1,0] \times R^2 \mid \text{there exists } (x,u) \in \mathcal{E} \text{ with } |x-y| + |u-w| < h\},$$

we have $\mathcal{B}_{2h}(\mathcal{E}) \subset \mathcal{N}$. For any given $(x,u_0) \in [-1,0] \times R^2$, the (exact) solution $u(t)$ of (3.9) with $u(t_0) = u_0$ enters $\mathcal{B}_h(\mathcal{E})$ at some finite time t, by Proposition 2.4. Define the time of first entry by

$$T(x,u_0) = \inf_{t>0}\{t \mid u(t) \in \mathcal{B}_h(\mathcal{E})\}.$$

It is well known, and easy to prove, that $T(x,u_0)$ is an upper semicontinuous function of its arguments. Since an upper semicontinuous function defined on a compact set attains its maximum, then by confining attention to those solutions of (3.9) satisfying (3.4) we obtain that

$$T_0 := \sup\{T(x,u_0) \mid \ -1 \le x \le 0,\ |u_0| \le M_u \ < \infty\}.$$

Now, there exists $\delta_0 > 0$, independent of x, u_0, such that for $\delta < \delta_0$, any δ-approximate solution w to (3.9) with $w(0) = u_0$, $|u_0| < M_u$, satisfies $|w(t) - u(t)| < h$ for $0 \le t \le T_0$ (to prove this, use Gronwall's inequality). It follows that $w(t) \in \mathcal{B}_{2h}(\mathcal{E}) \subset \mathcal{N}$ for some $t < T_0$, and this completes the proof. □

Our strategy for proving $\|S\|_{H^1} \to 0$ as $t \to \infty$ involves a bootstrapping argument, based on an estimate which we will derive from Proposition 3.1, namely,

$$\limsup_{t\to\infty} \|S(\cdot,t)\|_{H^1} \le K\alpha \limsup_{t\to\infty} \|G(\cdot,u(\cdot,t))\|_{L^2}. \tag{3.12}$$

The idea is that if α is small, then the right side of (3.12) is small, so S is small and $u(x,\cdot)$ is a δ-approximate solution of (3.9) with δ small. This should lead to

$G(x,u(x,t))$ becoming small for large t, so that the right side of (3.12) is smaller, and S is smaller, etc.

The proof of Theorem 3.5 in §6 depends crucially on the following characterization of δ-approximate solutions to (3.9). We say a point $p \in R^2$ is a *recurrent point* of a δ-approximate solution $u(t)$ of (3.9) if there exists a sequence $\{t_n\}_{n=1}^{\infty}$ with $t_n \to \infty$ as $n \to \infty$, such that

$$u(t_n) \to p \quad \text{as } n \to \infty.$$

DEFINITION 3.8. For $x \in [-1,0]$ we define the set $R(x,\delta)$ of possible recurrent points of δ-approximate solutions of (3.9) by

$$R(x,\delta) = \{p \in R^2 \mid p \text{ is a recurrent point of some } \delta\text{-approximate solution of (3.9)}\}.$$

PROPOSITION 3.9. *There is a constant K independent of x, such that for $\delta > 0$ sufficiently small,*

$$\sup_{p\in R(x,\delta)} |G(x,p)| \leq K\delta \quad \textit{for all } x \in [-1,0]. \tag{3.13}$$

This result is a quantitative improvement (for a flow with a parameter x) of a remark of Conley [2, p. 17] asserting without proof that for a strongly gradient-like flow, if the (persistent) error made in following solutions is smaller than ε for time 1, then asymptotically the approximate solution will lie in any given neighborhood of the set of rest points, if ε is small enough. Also see Norton [10], who studies the general structure of flows with persistent errors.

The proof of Proposition 3.9 is the most technical part of this paper, and is given in §§7–10. It requires exploiting the gradient-like structure of the system (Q), and decomposing the phase space of (3.9) into a nested sequence of positively invariant sets, in such a way that Proposition 3.9 can be deduced. Recalling that Conley [2] defined a "Morse decomposition" of a flow to be a nested sequence of (positively *and* negatively) invariant sets, we call the (one-sided) decomposition that we require a "semi-Morse decomposition."

DEFINITION 3.10. A *semi-Morse decomposition* for the system (Q) is a finite nested sequence $\{M_j\}_{j=0}^{i}$ of subsets of the phase space $[-1,0] \times R^2$, with $M_0 \supset M_1 \supset \cdots \supset M_i$, such that each M_j is positively invariant for solutions of (Q).

In such a decomposition, if a solution $u(t)$ leaves one of the sets $M_j \setminus M_{j+1}$, it can do so only by entering M_{j+1}. We remark that this definition is related to the concept of an "index filtration" introduced by Franzosa [3] to study Morse decompositions via the Conley index.

The use we make of a semi-Morse decomposition in proving Proposition 3.9 is as a quantitative, rather than theoretical, tool (cf. [2], [3]). The result we seek to prove is the following.

PROPOSITION 3.11. *There exist positive constants δ_0 and K, such that for any $\delta < \delta_0$ and $x \in [-1,0]$, there exist nested sets*

$$M_0(x,\delta) \supset M_1(x,\delta) \supset \cdots \supset M_i(x,\delta),$$

which are positively invariant for δ-approximate solutions of (3.9) *and are such that*

(a) $|u| \leq M_u$ *implies* $u \in M_0(x,\delta)$;

(b) *For any given $j = 0,\ldots,i$, either every δ-approximate solution of* (3.9) *that lies in $M_j(x,\delta)$ must enter $M_{j+1}(x,\delta)$, or*

$$|G(x,u)| \leq K\delta \quad \textit{for all } u \in M_j(x,\delta) \setminus M_{j+1}(x,\delta). \tag{3.14}$$

Given this result, it is clear that for $\delta < \delta_0$ and any $x \in [-1,0]$, if $p \in R^2$ is a recurrent point for some δ-approximate solution of (3.9), then for some j, the point $p \in M_j(x,\delta) \setminus M_{j+1}(x,\delta)$, which must be a set where (3.14) holds. This will finish the proof of Proposition 3.9. We note that since the bounds available for δ do not depend on x, it will be necessary to study in particular detail the flow of (Q) for values of x near to where $\overline{T} = \overline{T}_m$ or $\overline{T}_M$ and saddle-node bifurcations occur.

4. Bounds for small α. In this section we prove Proposition 3.1 and the estimates for (σ, Z) in Theorem 3.3. Our method of obtaining estimates on S is to use the parabolic smoothing properties of the heat operator, and the variation of constants formula.

On the Hilbert space $X = L^2(-1,0)$, let Λ denote the operator with domain

$$D(\Lambda) = \{S \in H^2(-1,0) \mid S_x(-1) = 0 = S(0)\}$$

given by $\Lambda S = -S_{xx}$. The operator Λ is self-adjoint and positive, and $\lambda := \pi^2/4$ is its first eigenvalue. Parabolic smoothing estimates, for equations of the form $S_t + \Lambda S = f(t,S)$ are well discussed in Henry's book [6]; here we only need the following estimate, which has an elementary proof via eigenfunction expansions (see [6, p. 17]). Define

$$H_b^1 = \{v \in H^1(-1,0) \mid v(0) = 0\}$$

and recall the elementary estimates

$$\|v\|_{L^2} \le \|v\|_{L^\infty} \le \|v_x\|_{L^2} \quad \text{for } v \in H_b^1.$$

For $v \in H_b^1$, we define $\|v\|_{H^1} := \|v_x\|_{L^2}$.

LEMMA 4.1. *For $t > 0$ we have*

$$\text{(4.1)} \qquad \begin{aligned} \|e^{-\Lambda t}v\|_{H^1} &\le \|v\|_{H^1}e^{-\lambda t}, & v &\in H_b^1, \\ \|e^{-\Lambda t}v\|_{H^1} &\le \|v\|_{L^2}K(t), & v &\in L^2, \end{aligned}$$

where

$$\text{(4.2)} \qquad K(t) = \begin{cases} (2et)^{-1/2} & \text{for } t < \frac{1}{2}\lambda, \\ \lambda^{1/2}e^{-\lambda t} & \text{for } t > \frac{1}{2}\lambda. \end{cases}$$

For later reference, we remark that

$$\text{(4.3)} \qquad \int_0^\infty K(t)\,dt = \frac{3}{\pi\sqrt{e}} < \frac{2}{3}.$$

Now, apply the variation of constants formula to the first equation in (3.1). Suppressing the x-dependence, we obtain

$$\text{(4.4)} \qquad S(t) = e^{-\varepsilon\Lambda t/\alpha}S_0 + \int_0^t e^{-\varepsilon\Lambda(t-\tau)/\alpha}(aS+b)(\tau)\,d\tau.$$

Taking the norm and using the estimates of Lemma 4.1, we get

$$\text{(4.5)} \quad \|S(t)\|_{H^1} \le e^{-\varepsilon\lambda t/\alpha}\|S_0\|_{H^1} + \int_0^t K\left(\frac{\varepsilon(t-\tau)}{\alpha}\right)(L\varepsilon^{-1}\|S(\tau)\|_{L^2} + \|b(\tau)\|_{L^2})\,d\tau.$$

To prove Proposition 3.1, define

$$M_S(t) = \sup_{0\le\tau\le t} \|S(\tau)\|_{H^1}\, e^{\varepsilon\lambda\tau/2\alpha} e^{-\varepsilon\lambda t/2\alpha}, \qquad M_b(t) = \sup_{0\le\tau\le t} \|b(\tau)\|_{L^2}.$$

Let $T > 0$. From (4.5) we obtain, for $t \le T$,

$$e^{\varepsilon\lambda t/2\alpha}\|S(t)\|_{H^1} \le \|S_0\|_{H^1} + e^{\varepsilon\lambda t/2\alpha}\left[\varepsilon^{-1}LM_S(t)\int_0^t K\left(\frac{\varepsilon(t-\tau)}{\alpha}\right)e^{\varepsilon\lambda(t-\tau)/2\alpha}\,d\tau \right.$$
$$\left. + M_b(t)\int_0^t K\left(\frac{\varepsilon(t-\tau)}{\alpha}\right)d\tau\right].$$

Using (4.3), and since $\int_0^\infty K(t)e^{\lambda t/2}\,dt = \tilde{K}$, the right side is less than

$$\|S_0\|_{H^1} + e^{\varepsilon\lambda T/2\alpha}\left[\alpha\varepsilon^{-2}L\tilde{K}M_S(T) + \frac{2}{3}\alpha\varepsilon^{-1}M_b(T)\right].$$

Hence

$$(1-\alpha\varepsilon^{-2}L\tilde{K})M_S(T) \le \|S_0\|_{H^1}e^{-\varepsilon\lambda T/2\alpha} + \frac{2\alpha}{3\varepsilon}M_b(T),$$

and Proposition 3.1 follows.

To prove the estimates for (σ, Z) in Theorem 3.3, let $\mathring{u} = (\mathring{\sigma}, \mathring{Z})$, so that $(0, \mathring{u})$ is the solution of (JSO$_1$) with $\alpha = 0$ having initial data $(0, \sigma_0, Z_0)$. We have

$$u_t = G(x,u) + HS, \qquad \mathring{u}_t = G(x,\mathring{u}).$$

Then $(u-\mathring{u})_t = G(x,u) - G(x,\mathring{u}) + HS$. Using the fact that $u(x,0) = \mathring{u}(x,0)$, we obtain for $0 \le t \le T$,

$$|u(x,t) - \mathring{u}(x,t)| \le \varepsilon^{-1}L\int_0^t (|u(x,\tau)-\mathring{u}(x,\tau)| + |S(x,\tau)|)\,d\tau.$$

By the estimates in Corollary 3.2, for some constant K independent of α, depending on T,

$$\int_0^T |S(x,\tau)|\,d\tau \le K\alpha, \qquad -1 \le x \le 0.$$

Then as $\alpha \to 0$, Gronwall's inequality implies $u(x,t) - \mathring{u}(x,t) = O(\alpha)$ on $[0,T]$, uniformly with respect to x on [-1,0].

5. Stability of discontinuous steady states. In this section we prove Theorem 3.4. Our first step is to obtain another bound for S. Write the perturbation in u as $w(x,t) = u(x,t) - \overline{u}(x)$, and define

$$M_w(t) = \sup_{\substack{0\le\tau\le t\\ x\in[-1,0]\setminus\mathcal{U}}} |w(x,t)|, \qquad M_S(t) = \sup_{0\le\tau\le t} \|S(\cdot,t)\|_{H^1}.$$

Using (3.4) and the fact that $b(x,\overline{u}(x)) = 0$, we have, for $0 \le \tau \le t$,

$$\|b(\cdot,u(\cdot,\tau))\|_{L^2}^2 \le \varepsilon^{-2}L^2\|w(\cdot,\tau)\|_{L^2}^2 \le \varepsilon^{-2}L^2(M_w(t)^2 + \delta^2 M_{\bar{u}}^2).$$

Assuming $\alpha\varepsilon^{-2}L\tilde{K} < \frac{1}{2}$ we may invoke Proposition 3.1 to conclude that

$$M_S(t) \le 2\delta e^{-\varepsilon\lambda t/2\alpha} + \tfrac{4}{3}\alpha\varepsilon^{-2}L(M_w(t) + \delta M_u). \tag{5.1}$$

Next, since $G(x, \overline{u}(x)) = 0$ we may write

$$G(x, u) = 0 + J(x)w + E(x, w),$$

where the Jacobian matrix

$$J(x) = \frac{\partial G}{\partial u}(x, \overline{u}) = \begin{pmatrix} -\frac{1}{\varepsilon}(\overline{Z} + 1 + \varepsilon) & \frac{1}{\varepsilon}(\overline{T} - \overline{\sigma}) \\ -\frac{1}{\varepsilon}(\overline{T} - 2\overline{\sigma}) & -1 \end{pmatrix}, \tag{5.2}$$

and for some constant K, $|E(x, w)| \le \varepsilon^{-1}K|w|^2$ for $x \in [-1, 0]$, all w. Here and below (in this section), K denotes a generic constant independent of α and ε.

From the hypotheses of the theorem and the results of [8], it follows that the matrix norm $|J(x)| \le K/\varepsilon$ for all x and the eigenvalues of $J(x)$ have real parts bounded above by $-\gamma < 0$, where γ and K depend on h_0 but not on α or ε. Then (replacing K by a larger K if necessary) we have that for each $x \in [-1, 0]$,

$$|e^{tJ(x)}| \le \varepsilon^{-1}Ke^{-\gamma t}. \tag{5.3}$$

For each x, the perturbation w satisfies

$$w_t = J(x)w + E(x, w) + H(x, u)S.$$

Applying the variation of constants formula, we have

$$w(x, t) = e^{tJ}w_0 + \int_0^t e^{(t-\tau)J}(E + HS)(\tau)\, d\tau,$$

which we may bound as follows for $x \in [-1, 0] \setminus \mathcal{U}$:

$$|w(x, t)| \le \varepsilon^{-1}K\delta + \varepsilon^{-1}K\int_0^t e^{-\gamma(t-\tau)}(\varepsilon^{-1}K|w(x, \tau)|^2 + \varepsilon^{-1}L|S(x, \tau)|)\, d\tau.$$

Using estimate (5.1) and the fact that

$$\int_0^t e^{-\gamma(t-\tau)}|w(x, \tau)|^2\, d\tau \le \gamma^{-1}M_w(t)^2,$$

we find, taking the supremum over $x \in [-1, 0] \setminus \mathcal{U}$ (and taking K larger if necessary), that

$$M_w(t) \le K\delta(\varepsilon^{-1} + \alpha\varepsilon^{-4}) + K\alpha\varepsilon^{-4}M_w(t) + K\varepsilon^{-2}M_w(t)^2, \tag{5.4}$$

or

$$0 \le K_0\delta - K_1M_w(t) + K_2M_w(t)^2, \tag{5.5}$$

where

$$K_0 = K(\varepsilon^{-1} + \alpha\varepsilon^{-4}), \quad K_1 = 1 - K\alpha\varepsilon^{-4}, \quad K_2 = \varepsilon^{-2}K.$$

Suppose α is so small that $K\alpha\varepsilon^{-4} < \frac{1}{2}$, and δ is sufficiently small. Then, since $M_w(t)$ is continuous in t and $M_w(0) < \delta$, and since (5.5) is violated if $M_w(t) = 4K_0\delta/K_1$, we find

$$M_w(t) < 8K_0\delta \quad \text{for all } t > 0.$$

It then follows from (5.1) that for some K,

$$M_S(t) < K\delta \quad \text{for all } t > 0$$

as well. This finishes the proof of Theorem 3.4.

6. Convergence as $t \to \infty$. Here we prove Theorem 3.5, assuming the validity of Proposition 3.9, whose proof appears in §§7–10. In this section, K denotes a generic constant independent of α (but K may depend on ε and M).

Start with estimate (3.5) in Proposition 3.1, and translate from the interval $[0,t]$ to $[t_0, t_0+t]$ for $t_0 > 0$ to obtain for $t > 0$,

$$\|S(\cdot, t_0 + t)\|_{H^1} \le 2\|S(\cdot, t_0)\|_{H^1} e^{-\varepsilon\lambda t/2\alpha} + \frac{4\alpha}{3\varepsilon} \sup_{0\le\tau\le t} \|b(\cdot, u(\cdot, t_0+\tau))\|_{L^2}.$$

Letting $t \to \infty$, then $t_0 \to \infty$, we conclude that

$$S_\infty := \limsup_{t\to\infty} \|S(\cdot,t)\|_{H^1} \le \frac{4\alpha}{3\varepsilon} \limsup_{t\to\infty} \|b(\cdot, u(\cdot,t))\|_{L^2}. \tag{6.1}$$

Since $b = G_1$, this yields (3.12).

Our goal is to show that $S_\infty = 0$, then establish that $\lim_{t\to\infty} u(x,t)$ exists for each x by studying the ODE for u in (3.1). Recall that $b(x, u(x,t))$ is uniformly bounded independent of x, t, and α. Applying the Lebesgue dominated convergence theorem, we may easily infer that

$$\limsup_{t\to\infty} \|b(\cdot, u(\cdot,t))\|_{L^2} \le \| \limsup_{t\to\infty} | b(\cdot, u(\cdot,t))| \|_{L^2}. \tag{6.2}$$

We know that $|b(x,u)| \le |G(x,u)|$, so that b may be replaced by G on the right-hand side of (6.2). For large t we have the bound

$$|H(x, u(x,t))S(x,t)| \le K\|S(\cdot,t)\|_{H^1} \le 2KS_\infty.$$

Thus for x fixed and t large, we may consider $u(x,t)$ as a δ-approximate solution of (3.9) with

$$\delta = 2KS_\infty. \tag{6.3}$$

With $R(x,\delta)$ defined (see Definition 3.8) as the set of possible recurrent points of such δ-approximate solutions, it is clear that

$$\lim_{t\to\infty} \operatorname{dist}(u(x,t), R(x,\delta)) = 0, \tag{6.4}$$

and hence, for all $x \in [-1,0]$,

$$\limsup_{t\to\infty} |G(x,u(x,t))| \le \sup_{p\in R(x,\delta)} |G(x,p)|. \tag{6.5}$$

Invoking Proposition 3.9 and (6.1), (6.2), we obtain

$$S_\infty \le K\alpha\delta \le 2K^2\alpha S_\infty.$$

Hence if α is sufficiently small, so that $2K^2\alpha < 1$, it follows that $S_\infty = 0$.

To show that $\lim_{t\to\infty} u(x,t)$ exists, observe that since $S_\infty = 0$, any recurrent point of $u(x,\cdot)$ lies in $R(x,\delta)$ for all $\delta > 0$. By Proposition 3.9, such recurrent points must be critical points of (Q). Since critical points of (Q) are isolated and $u(x,t)$ is bounded, it follows that $u(x,t)$ must converge to some single critical point as $t \to \infty$. This finishes the proof of Theorem 3.5, assuming the validity of Proposition 3.9.

7. Local coordinates and estimates near the manifold of equilibria.

The next three sections are devoted to the proof of Proposition 3.11, from which Proposition 3.9 follows. The construction of the nested sets $M_j(x,\delta)$, which are positively invariant for δ-approximate solutions of (3.9) with x fixed, must be done differently in three cases. For some small $h_0 > 0$, which will be fixed later, the cases correspond to:

(i) $\overline{T} \le \overline{T}_m - h_0$ or $\overline{T} \ge \overline{T}_M + h_0$. In this case system (Q) has a single uniformly attracting critical point, A or C;

(ii) $\overline{T}_m + h_0 \le \overline{T} \le \overline{T}_M - h_0$. System (Q) has three uniformly hyperbolic critical points, a saddle B, and two attracting points A and C;

(iii) $|\overline{T} - \overline{T}_m| < h_0$ or $|\overline{T} - \overline{T}_M| < h_0$. System (Q) undergoes a saddle-node bifurcation.

To define the sets $M_j(x,\delta)$, we need to construct local coordinate systems near the manifold $\mathcal{E}$ of critical points defined in (3.10), using linear approximation of (Q) and the unstable and center manifold theorems. The number h_0 above, for example, is determined by applying the center manifold theorem in a neighborhood of the saddle-node bifurcation points.

In the rest of this section, we study the manifold of critical points $\mathcal{E}$ and describe local coordinate systems for (x,u) near the manifold $\mathcal{E}$ in three regimes:

(a) near the curves of stable critical points $(x,A(x))$ and $(x,C(x))$;

(b) near the curve of saddle points $(x,B(x))$;

(c) near the saddle-node bifurcation points.

In each coordinate system, we derive the basic estimates for δ-approximate solutions of (3.9) that will be used to construct the semi-Morse decomposition. Some of the sets $M_j(x,\delta)$ will not depend on δ; in such cases, the second argument may be omitted.

7.1. Critical points of (3.9).

Without loss of generality, we may assume that $\overline{f} > \overline{T}_M$, so that as x ranges from zero to -1, $\overline{T} = -\overline{f}x$ ranges from zero to $\overline{f} > \overline{T}_M$. Then we may define $x_M < x_m$ in [-1,0] so that

$$-\overline{f}x_M = \overline{T}_M, \qquad -\overline{f}x_m = \overline{T}_m. \tag{7.1}$$

From (1.2)–(1.4), we have that the manifold of critical points $\mathcal{E}$ is the union of three curves:

$$\begin{aligned}\mathcal{E} = \{(x,A(x)) \mid x_M \le x \le 0\} \;&\cup\; \{(x,B(x)) \mid x_M \le x \le x_m\} \\ &\cup\; \{(x,C(x)) \mid -1 \le x \le x_m\}.\end{aligned}$$

For $P = A$, B, or C, the critical point $P(x)$ has the form

$$P(x) = (\sigma_P(x), Z_P(x)) = \left(\frac{\xi}{1+\xi^2}, \frac{-\xi^2}{1+\xi^2}\right), \tag{7.2}$$

where $\xi = \xi_P(x)$ is a root of $\omega(\xi) = \overline{T}$. Here

$\xi_A(x)$ is the smallest root, defined for $x_M \le x \le 0$;
$\xi_B(x)$ is the middle root, defined for $x_M \le x \le x_m$;
$\xi_C(x)$ is the largest root, defined for $-1 \le x \le x_m$.

Let us recall now from [8] some properties of the linearization of (3.9) at the critical points $(x, P(x))$. For $P = A$, B, or C define

$$J_P(x) = \frac{\partial G}{\partial u}(x, P(x)) = \begin{pmatrix} -\frac{1}{\varepsilon}\left(\varepsilon + \frac{1}{1+\xi^2}\right) & \xi \\ -\frac{1}{\varepsilon}\left(\varepsilon\xi - \frac{\xi}{1+\xi^2}\right) & -1 \end{pmatrix}. \tag{7.3}$$

We have $\varepsilon \operatorname{tr} J_P(x) = -2\varepsilon - 1/(1+\xi_P^2) < 0$ always. Also, $\varepsilon \det J_P(x) = (1+\xi_P^2)\omega'(\xi_P)$, which is positive for $P = A$, $x_M < x \le 0$, and for $P = C$, $-1 \le x < x_m$, and is negative for $P = B$, $x_M < x < x_m$. Correspondingly, we have the following (see [8]).

LEMMA 7.1.

(1) *For $x_M < x \le 0$, $J_A(x)$ has two negative eigenvalues.*
(2) *For $-1 \le x < x_m$, $J_C(x)$ has two eigenvalues with negative real part.*
(3) *For $x_M < x < x_m$, $J_B(x)$ has one negative and one positive eigenvalue.*
(4) *For $x = x_M$, $J_A(x) = J_B(x)$ has one zero and one negative eigenvalue.*
(5) *For $x = x_m$, $J_C(x) = J_B(x)$ has one zero and one negative eigenvalue.*

Note also since $\varepsilon\xi \le \omega(\xi) = \overline{T}$, we have $\varepsilon|J_P(x)| \le K$ independent of x, P, and ε, where $|\cdot|$ is the matrix norm. For ε fixed, $|J_P(x)| \le K_0$ independent of x and P.

7.2. Local coordinates. For later reference, we note that if $(x, P(x))$ is a critical point of (3.9), then since $G(x, P(x)) = 0$ and G is quadratic, we may write

$$G(x,u) = J_P(x)(u - P(x)) + E(x, u - P(x)), \quad \text{where } |E(x,w)| \le K_1|w|^2 \tag{7.4}$$

for all $w \in R^2$, and some constant K_1 independent of x.

Case (a). We study uniformly stable branches of the manifold of critical points. For $h > 0$ given, put

$$\mathcal{E}_a(h) = \{(x, A(x)) \mid x_M + h \le x \le 0\} \cup \{(x, C(x)) \mid -1 \le x \le x_m - h\}. \tag{7.5}$$

LEMMA 7.2. *Let $h > 0$. Then there exist positive constants $K_2 > 0$ and $\gamma > 0$ such that whenever $(x, P(x)) \in \mathcal{E}_a(h)$, then with $J = J_P(x)$ we have*

$$|\exp(Jt)| \le K_2 e^{-\gamma t}, \qquad t \ge 0, \tag{7.6}$$

and, with $D = D_P(x)$ defined by

$$D = \int_0^\infty \exp(Jt)^T \exp(Jt)\, dt, \tag{7.7}$$

we have $D^T = D$, $J^T D + DJ^T = -I$, *where* I *is the* 2×2 *identity matrix, and for all* $w \in R^2$,

$$w^T DJw = -\frac{1}{2}|w|^2, \qquad \frac{1}{2|J|}|w|^2 \le w^T Dw \le \frac{K_2^2}{2\gamma}|w|^2. \tag{7.8}$$

For the proof, see [1, Appendix 4].

Now suppose that $u = u(x,t)$ is a δ-approximate solution of (3.9), with x fixed so that $(x, P(x)) \in \mathcal{E}_a(h)$ as above. That is,

$$u_t = G(x,u) + f(t), \quad \text{where } |f(t)| \le \delta, \tag{7.9}$$

and f is continuous. Define $w(t) = u(x,t) - P(x)$. Using (7.4) we have, with $J = J_P(x)$,

$$w_t = Jw + E(x,w) + f(t), \qquad |f(t)| \le \delta.$$

From this we derive, using the inequalities (7.8),

$$\frac{d}{dt} w^T Dw = 2w^T D(Jw + E + f) \le -|w|^2 + \frac{K_2^2}{\gamma}|w|(K_1|w|^2 + \delta).$$

Thus we obtain

$$\frac{d}{dt} w^T Dw \le -\frac{1}{2}|w|^2 < 0, \tag{7.10}$$

provided

$$|w|\frac{K_1 K_2^2}{\gamma} \le \frac{1}{4} \quad \text{and} \quad \frac{|w|}{4} > \left(\frac{K_2^2}{\gamma}\right)\delta.$$

Using (7.8), we conclude that (7.10) also follows if

$$K_a \delta \le (w^T Dw)^{1/2} \le \frac{1}{K_a}, \tag{7.11}$$

where K_a is a suitable constant (depending on h but not on x).

Case (b). We consider a uniformly hyperbolic branch of saddle points in the manifold of critical points. For $h > 0$ given, put

$$\mathcal{E}_b(h) = \{(x, B(x)) \mid x_M + h \le x \le x_m - h\}. \tag{7.12}$$

Denote the eigenvalues of $J_B(x)$ by $\lambda_1(x)$, $\lambda_2(x)$, with $\lambda_1(x) < 0 < \lambda_2(x)$ for $x_M < x < x_m$. Let $R(x)$ be a smooth matrix defined for $x \in (x_M, x_m)$ whose columns $r_1(x)$, $r_2(x)$ are eigenvectors of $J_B(x)$ corresponding, respectively, to $\lambda_1(x)$ and $\lambda_2(x)$. For $(x,u) \in (x_M, x_m) \times R^2$, change variables to $(x, \tilde{w})$ by requiring

$$u - B(x) = R(x)\tilde{w} = r_1(x)\tilde{w}_1 + r_2(x)\tilde{w}_2. \tag{7.13}$$

Invoking the unstable manifold theorem, see [7], we have the following.

LEMMA 7.3. *Let* $h > 0$. *Then there exists* $\eta_0 > 0$ *such that, for* $x_M + h \le x \le x_m - h$, *the unstable manifold of* $B(x)$ *in* (3.9) *is parametrized locally by*

$$\tilde{w}_1 = \phi(x, \tilde{w}_2) \quad \textit{for } |\tilde{w}_2| \le \eta_0, \tag{7.14}$$

where $\phi(x,y)$ is smooth, with $\phi(x,0) = 0 = \partial\phi/\partial y(x,0)$ for all x.

We make a further change of variables to (x,w), putting

$$w_1 = \tilde{w}_1 - \phi(x,\tilde{w}_2), \qquad w_2 = \tilde{w}_2. \tag{7.15}$$

Then, given $h > 0$, there exists $\eta_1 > 0$ such that the change of variables from (x,u) to (x,w) is smooth and defined for

$$x_M + h \le x \le x_m - h, \qquad |w|_\infty = \max(|w_1|,|w_2|) \le \eta_1. \tag{7.16}$$

In terms of the variables (x,w), the system (3.9) may be written in the form

$$\begin{aligned} w_{1t} &= \lambda_1(x)w_1 + F_1(x,w_1,w_2), \\ w_{2t} &= \lambda_2(x)w_2 + F_2(x,w_1,w_2), \end{aligned} \tag{7.17}$$

where, with $F(x,w) = (F_1,F_2)$, F is smooth, with $F(x,0) = 0 = \partial F/\partial w(x,0)$, and moreover, since $w_1 = 0$ is an invariant manifold of (7.17), $F_1(x,0,w_2) = 0$ for $|w_2| \le \eta_1$. Hence for some constant K_0, we have the estimates

$$|F_1(x,w)| \le K_0|w_1||w|_\infty, \qquad |F_2(x,w)| \le K_0|w|_\infty^2 \quad \text{for } |w|_\infty \le \eta_1. \tag{7.18}$$

Suppose now $u = u(x,t)$ is a δ-approximate solution of (3.9), where $(x,B(x)) \in \mathcal{E}_b(h)$ as above. For some $\eta > 0$ (depending on h but not on x), if $|u(x,t) - B(x)| < \eta$, we may change variables from (x,u) to (x,w) as above, obtaining

$$\begin{aligned} w_{1t} &= \lambda_1(x)w_1 + F_1(x,w_1,w_2) + f_1(t), \\ w_{2t} &= \lambda_2(x)w_2 + F_2(x,w_1,w_2) + f_2(t), \end{aligned} \tag{7.19}$$

where for some $K > 0$ (depending on h but not on x), $|f(t)|_\infty \le K\delta$. Then so long as $|w|_\infty \le \eta_1$, we have

$$\begin{aligned} \frac{d}{dt}w_1^2 = 2w_1(\lambda_1 w_1 + F_1 + f_1) &\le 2\lambda_1 w_1^2 + 2|w_1|(K_0|w_1||w|_\infty + K\delta) \\ &\le \lambda_1 w_1^2 < 0, \end{aligned} \tag{7.20}$$

provided that

$$2K_0|w|_\infty < \tfrac{1}{2}|\lambda_1| \quad \text{and} \quad 4K\delta < |\lambda_1||w_1|. \tag{7.21}$$

Similarly,

$$\begin{aligned} \frac{d}{dt}w_2^2 = 2w_2(\lambda_2 w_2 + F_2 + f_2) &\ge 2\lambda_2 w_2^2 - 2|w_2|(K_0|w|_\infty^2 + K\delta) \\ &\ge \lambda_2 w_2^2 > 0, \end{aligned} \tag{7.22}$$

provided that

$$2K_0|w_2| < \tfrac{1}{2}|\lambda_2| \quad \text{and} \quad 2K_0|w_1|^2 < \tfrac{1}{2}\lambda_2|w_2| \quad \text{and} \quad 2K\delta < \tfrac{1}{2}|\lambda_2||w_2|. \tag{7.23}$$

Case (c). We study the branches of critical points close to the bifurcation points $(x_M, A(x_M))$ and $(x_m, C(x_m))$. For $h > 0$, define

$$\begin{aligned} \mathcal{E}_c(h) =& \{(x,P(x)) \mid P = A \text{ or } B \text{ and } x_M \le x \le x_M + h\} \\ &\cup \ \{(x,P(x)) \mid P = C \text{ or } B \text{ and } x_m - h \le x \le x_m\}. \end{aligned} \tag{7.24}$$

Note that $\mathcal{E} = \mathcal{E}_a(h) \cup \mathcal{E}_b(h) \cup \mathcal{E}_c(h)$ for any $h > 0$.

We study a neighborhood of the point (x_0, P_0), which denotes either saddle-node bifurcation point $(x_M, A(x_M))$ or $(x_m, C(x_m))$. From Lemma 7.1, The Jacobian matrix $J_0 = \partial G/\partial u(x_0, P_0)$ at this point has eigenvalues $\lambda_1 < 0$, $\lambda_2 = 0$. We let r_1, r_2 denote corresponding right eigenvectors of J_0, and let R be the matrix with columns r_1, r_2.

Now, we wish to augment system (3.9) by regarding x as an unknown satisfying the equation $\dot{x} = 0$. Define

$$U = \begin{pmatrix} u \\ x \end{pmatrix}, \qquad \tilde{G}(U) = \begin{pmatrix} G(x,u) \\ 0 \end{pmatrix}, \qquad \tilde{J} = \frac{\partial \tilde{G}}{\partial U}(U_0) = \begin{pmatrix} J_0 & G_x \\ 0 & 0 \end{pmatrix},$$

where $U_0 = (P_0, x_0)$ and $G_x = \partial G/\partial x(x_0, P_0)$. System (3.9) becomes the augmented system $U_t = \tilde{G}(U)$. The Jacobian matrix of this system at U_0 is $\tilde{J}$, which has (generalized) eigenvectors of the form

$$\tilde{r}_1 = \begin{pmatrix} r_1 \\ 0 \end{pmatrix}, \qquad \tilde{r}_2 = \begin{pmatrix} r_2 \\ 0 \end{pmatrix}, \qquad \tilde{r}_3 = \begin{pmatrix} a_1 r_1 \\ 1 \end{pmatrix},$$

satisfying $\tilde{J}\tilde{r}_1 = \lambda_1 \tilde{r}_1$, $\tilde{J}\tilde{r}_2 = 0$, $\tilde{J}\tilde{r}_3 = a_2 \tilde{r}_2$, where a_1, a_2 are determined by requiring

$$G_x = -\lambda_1 a_1 r_1 + a_2 r_2. \tag{7.25}$$

Thus, if $\tilde{R}$ is the matrix with columns $\tilde{r}_1$, $\tilde{r}_2$, $\tilde{r}_3$, we have

$$\tilde{J}\tilde{R} = \tilde{R}\Lambda, \qquad \Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & 0 & a_2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Next, make a change of variables from $U = (u, x)$ to $\tilde{W} = (\tilde{w}, y)$, where

$$U - U_0 = \tilde{R}\tilde{W}, \quad \text{i.e.,} \quad u - P_0 = r_1\tilde{w}_1 + r_2\tilde{w}_2 + a_1 r_1 y, \quad x - x_0 = y.$$

In these variables, the system becomes

$$\tilde{W}_t = \Lambda\tilde{W} + \tilde{F}(\tilde{W}), \tag{7.26}$$

where $\tilde{F}(0) = 0$ and $\partial\tilde{F}/\partial\tilde{W}(0) = 0$, and in fact, $\tilde{F}_3 = 0$.

We may now apply the center manifold theorem to system (7.26); see [7]. The result is the following.

LEMMA 7.4. *There exists $h_1 > 0$, and a neighborhood $\mathcal{N}_h$ of $\tilde{W} = 0$, such that* (7.26) *has a locally invariant manifold in $\mathcal{N}_h$ of the form*

$$\tilde{w}_1 = \phi(\tilde{w}_2, y), \quad \textit{with } \phi(0,0) = 0, \quad \partial\phi(0,0) = 0,$$

where $\partial\phi$ is the gradient of ϕ. Here ϕ is defined for $\max(|\tilde{w}_2|, |y|) < h_1$ *and ϕ is C^k (k is arbitrary, but $\mathcal{N}_h$ depends on k).*

The center manifold need not be unique, but since $\lambda_1 < 0$, the center manifold is a center-unstable manifold, so any solution $\tilde{W}(t)$ that lies in $\mathcal{N}_h$ for all $t < 0$ must lie in each such center-unstable manifold [7]. In particular, equilibria in $\mathcal{N}_h$ and their local unstable manifolds lie in the center manifold above.

We make a further change of variables, defining

$$w_1 = \tilde{w}_1 - \phi(\tilde{w}_2, y), \qquad w_2 = \tilde{w}_2. \tag{7.27}$$

For some $h_2 > 0$, the change of variables from $U = (u, x)$ to $W = (w, y)$ is C^k (for any fixed k) and is defined for $|W|_\infty = \max(|w_1|, |w_2|, |y|) \leq h_2$. In these variables, the system becomes

$$\begin{aligned} w_{1t} &= \lambda_1 w_1 + F_1(w_1, w_2, y), \\ w_{2t} &= a_2 y + F_2(w_1, w_2, y), \\ y_t &= 0, \end{aligned} \tag{7.28}$$

or $W_t = \Lambda W + F(W)$, where

$$F(0) = 0, \qquad \frac{\partial F}{\partial W}(0) = 0, \quad \text{and} \quad F_1(0, w_2, y) = 0; \tag{7.29}$$

the center manifold $w_1 = 0$ is locally invariant. For some $K_0 > 0$ we have the estimates

$$|F_1(w)| \leq K_0 |w_1| |W|_\infty, \qquad |F_2(W)| \leq K_0 |W|_\infty^2 \quad \text{for } |W|_\infty \leq h_2. \tag{7.30}$$

In what follows, we need to study a bit more closely the equation on the center manifold. From the results so far, there exists a C^k function $a(w_2, y)$ such that on the center manifold $w_1 = 0$ we have

$$w_{2t} = a(w_2, y)y + bw_2^2 = a_2 y + F_2(0, w_2, y), \tag{7.31}$$

where $a(0,0) = a_2$ and $b = \frac{1}{2}\partial^2 F_2/\partial w_2^2(0)$.

LEMMA 7.5. *With an appropriate choice of the null eigenvector r_2, we have that $b = 1$. Moreover, we have*

$$a_2 < 0 \quad at\ (x_M, A(x_M)), \qquad a_2 > 0 \quad at\ (x_m, C(x_m)).$$

Proof. Provided we show $b \neq 0$, we can arrange $b = 1$ by replacing r_2 by r_2/b. (The corresponding value of a_2 is replaced by ba_2.) To compute b and a_2, let ℓ be a left null eigenvector of J_0, satisfying $\ell J_0 = 0$. Then a_2 is determined by $a_2 = \ell G_x / \ell r_2$. From (7.3) we see that we may take, with $\xi = \xi_P(x_0)$,

$$r_2 = \begin{pmatrix} \varepsilon\xi \\ \varepsilon + \frac{1}{1+\xi^2} \end{pmatrix}, \qquad \ell = (1, \xi). \tag{7.32}$$

Then from (3.2) we compute

$$G_x = -\frac{\overline{f}}{\varepsilon}\begin{pmatrix} Z_P + 1 \\ -\sigma_P \end{pmatrix} = -\frac{\overline{f}}{\varepsilon}\begin{pmatrix} \frac{1}{1+\xi^2} \\ -\frac{\xi}{1+\xi^2} \end{pmatrix}.$$

Hence $\ell r_2 = \varepsilon\xi + \omega(\xi) > 0$, and $\ell G_x = (-\overline{f}/\varepsilon)(1-\xi^2)/(1+\xi^2) = \overline{f}(1+\xi^2) > 0$, which follows from $0 = \omega'(\xi) = \varepsilon + (1-\xi^2)/(1+\xi^2)^2$ at $\xi = \xi_P(x_0)$. So, with the choice of r_2 in (7.32), we have $a_2 > 0$ at both saddle-node bifurcation points.

To compute b, we use (7.29) and the fact that $\partial\phi(0,0)=0$ from Lemma 7.4. We have

$$2b = \frac{\partial^2 F_2}{\partial w_2^2}(0) = \frac{\partial^2 \tilde{F}_2}{\partial \tilde{w}_2^2}(0) = \ell\, \partial^2 G(r_2, r_2),$$

where $\partial^2 G$ is the Hessian of G. From (3.2) and (7.32) we compute $b = \frac{1}{2}\ell\, \partial^2 G(r_2,r_2) = \xi(\varepsilon(\xi^2-1) - 1/(1+\xi^2))$. Since $\omega'(\xi)=0$ we have $\varepsilon = (\xi^2-1)/(1+\xi^2)^2$, so we find $b = \xi^3(\xi^2-3)/(1+\xi^2)^2$. Now $\omega'(\sqrt{3}) = \varepsilon - \frac{1}{8} < 0$, so $\xi_A(x_M)^2 < 3 < \xi_C(x_m)^2$. Thus we find that $b<0$ at $(x_M, A(x_M))$, and $b>0$ at $(x_m, C(x_m))$. Replacing r_2 by r_2/b as discussed previously, the Lemma follows. □

Now suppose that $u = u(x,t)$ is a δ-approximate solution of (3.9), lying in a neighborhood of one of the saddle-node bifurcation points, where the change of variables from (u,x) to (w,y) is defined. In the (w,y) variables, we have

$$\begin{aligned} w_{1t} &= \lambda_1 w_1 + F_1(w_1,w_2,y) + f_1(t), \\ w_{2t} &= w_2^2 + a(w_2,y)y + g_2(w_1,w_2,y) + f_2(t), \\ y_t &= 0, \end{aligned} \tag{7.33}$$

where for some constant $K>0$, $|f(t)|_\infty \le K\delta$, and $|g_2(w_1,w_2,y)| \le K|w_1||W|_\infty$. Then, so long as $|W|_\infty \le h_2$, we have the following estimates:

$$|w_1|_t = (\operatorname{sgn} w_1)(\lambda_1 w_1 + F_1 + f_1) \le \lambda_1|w_1| + K_0|w_1||W|_\infty + K\delta \le \tfrac{1}{2}\lambda_1|w_1| < 0, \tag{7.34}$$

provided that

$$K_0|W|_\infty \le \tfrac{1}{4}|\lambda_1| \quad \text{and} \quad 4K\delta \le |\lambda_1||w_1|. \tag{7.35}$$

Also, we have

$$\operatorname{sgn}(w_{2t}) = \operatorname{sgn}(w_2^2 + ay), \tag{7.36}$$

provided

$$2K\max(\delta, h_2|w_1|) \le |w_2^2 + ay|. \tag{7.37}$$

8. A semi-Morse decomposition: case (i).

In this section we construct nested, positively invariant sets

$$M_0(x) \supset M_1(x) \supset M_2(x,\delta) \tag{8.1}$$

in the case (i) mentioned at the start of §7, namely, the case when the system (3.9) has a single uniformly attracting critical point. This is the simplest case.

We first mention that in all the cases (i), (ii), and (iii), the first positively invariant set $M_0(x)$ will be chosen, independently of both x and δ, in the form

$$M_0(x) = \{u = (\sigma, Z) \mid \sigma^2 + (Z+1)^2 \le K\} \tag{8.2}$$

for $K>1$ sufficiently large, so that $|u| \le M_u$ implies $u \in M_0(x)$. For a δ-approximate solution of (3.9), we may obtain

$$\frac{d}{dt}(\sigma^2 + (Z+1)^2) \le -2(\sigma^2 + (Z+1)Z) + (\sigma^2 + (Z+1)^2)^{1/2}\delta.$$

It follows that the set in (8.2) is positively invariant for δ-approximate solutions , if δ is sufficiently small, independent of x.

Now let $h > 0$. (This number will be fixed in §10.) We employ the variables introduced in §7, Case (a). For $(x, P(x)) \in \mathcal{E}_a(h)$, $P = A$ or C, we select $D = D_P(x)$ and put

$$N_P(x,\delta) = \{u = P(x) + w \mid (w^T D w)^{1/2} \leq K_a \delta\}, \tag{8.3}$$

where K_a is the constant in (7.11). According to (7.10) we have the following.

LEMMA 8.1. *The sets $N_P(x,\tilde{\delta})$ are positively invariant for δ-approximate solutions of* (3.9) *provided $\delta \leq \tilde{\delta} \leq 1/K_a^2$.*

Case (i) corresponds, by definition, to the cases

$$x_m + h \leq x \leq 0 \quad \text{and} \quad -1 \leq x \leq x_M - h. \tag{8.4}$$

Corresponding to these two subcases, the unique critical point of (3.9) is $P = A$ or C, respectively. For x satisfying (8.4), we put

$$M_1(x) = N_P\left(x, \frac{1}{K_a^2}\right), \qquad M_2(x,\delta) = N_P(x,\delta), \tag{8.5}$$

assuming $\delta \leq 1/K_a^2$. With these definitions, the sets in (8.1) now satisfy the following.

PROPOSITION 8.2. *Let $h > 0$. Then there exist constants $\delta_0 > 0$, $K > 0$ such that, if $\delta \leq \delta_0$ and x satisfies* (8.4), *then every δ-approximate solution of* (3.9) *that lies in $M_j(x,\delta)$ must enter $M_{j+1}(x,\delta)$, for $j = 0$ and 1. Hence, every recurrent point of such a δ-approximate solution must lie in $M_2(x,\delta)$. Moreover, we have*

$$\sup_{u \in M_2(x,\delta)} |G(x,u)| \leq K\delta. \tag{8.6}$$

Proof. If $u \in M_2(x,\delta)$, then by using Lemma 7.2 we have $|u - P(x)|^2 \leq 2K_0 K_a^2 \delta^2$, so (8.6) follows. From (7.10), (7.11) it is clear that any δ-approximate solution lying in $M_1(x)$ must enter $M_2(x,\delta)$. To prove the analogous assertion for $M_0(x)$, we apply Proposition 3.7. Let

$$\mathcal{N} = \{(x,u) \in [-1,0] \times R^2 \mid u \in M_1(x) \text{ if } x \text{ satisfies (8.4)}\}.$$

Now $\mathcal{N}$ is a neighborhood of $\mathcal{E}$, the manifold of critical points; so by Proposition 3.7, there exists $\delta_0 > 0$ such that, if $\delta < \delta_0$ and x satisfies (8.4), then any δ-approximate solution of (3.9) lying in $M_0(x)$ will enter $M_1(x)$. This finishes the proof of Proposition 8.2. □

9. A semi-Morse decomposition: case (ii). Case (ii) corresponds to

$$x_M + h \leq x \leq x_m - h. \tag{9.1}$$

Here h will be fixed in §10, as we have mentioned earlier. Because of the presence of the saddle B in this case, the semi-Morse decomposition we require has more sets than in the last section, seven in all: $M_0(x,\delta), \ldots, M_6(x,\delta)$. A sketch of the decomposition is given in Fig. 6; a detail of the neighborhood of the saddle B appears in Fig. 7. In

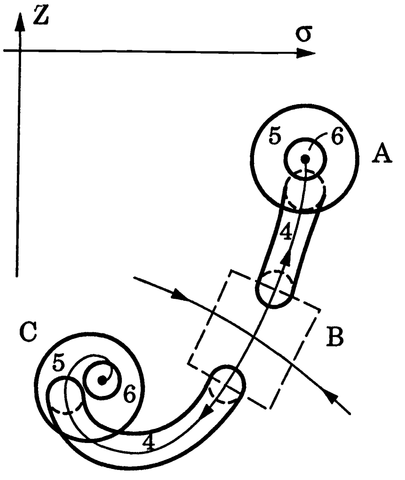


FIG. 6. *Semi-Morse decomposition, case* (ii).

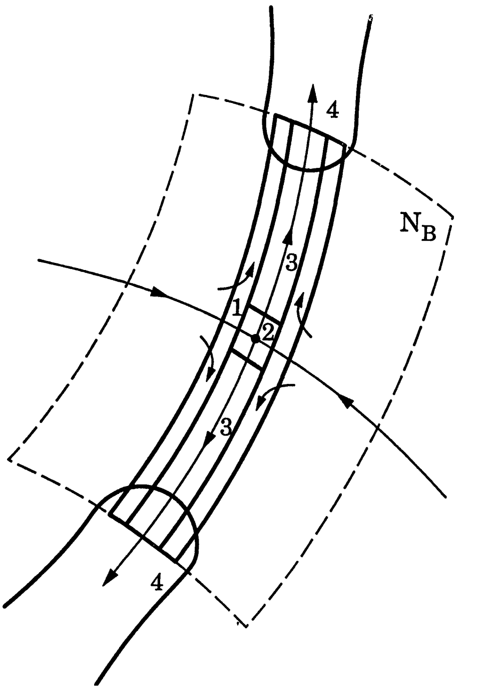


FIG. 7. *Detail near the saddle.*

these figures, the numbers $j = 0, \ldots, 6$ indicate the location of the region $M_j \setminus M_{j+1}$; the flow proceeds in increasing numerical order.

The smallest sets in the semi-Morse decomposition will be neighborhoods of the stable critical points A and C. Using Lemma 8.1 and the definition in (8.3), we will put

$$\begin{aligned} M_6(x,\delta) &= N_A(x,\delta) \cup N_C(x,\delta), \\ M_5(x) &= N_A\left(x, \frac{1}{K_a^2}\right) \cup N_C\left(x, \frac{1}{K_a^2}\right). \end{aligned} \tag{9.2}$$

We note that from (7.10), (7.11) it is clear that for $\delta < 1/K_a^2$, any δ-approximate solution of (3.9) lying in $M_5(x)$ must enter $M_6(x,\delta)$.

Now, to study the flow near the saddle point B, recall the results of §7, case (b). In the notation of that section, for x satisfying (9.1), and $\eta \le \eta_1$ (η_1 is given in (7.16)), we may define a neighborhood of $B(x)$ by

$$N_B(x,\eta) = \{u = B(x) + R(x)(w_1 + \phi(x,w_2), w_2)^T \mid |w|_\infty \le \eta\}. \tag{9.3}$$

From the estimates (7.20)–(7.23) we may infer the following.

LEMMA 9.1. *There exist constants $\eta_2, K_b > 0$ (depending on h but not on x) such that if $K_b\delta \le \eta_2$ and $u = u(x,t)$ is a δ-approximate solution lying in $N_B(x,\eta_2)$, then*

$$\begin{array}{ll} |w_1| \ge K_b\delta & \textit{implies } |w_1| \textit{ is strictly decreasing;} \\ |w_2| \ge K_b \max(\delta, w_1^2) & \textit{implies } |w_2| \textit{ is strictly increasing.} \end{array}$$

To define the next set $M_4(x,\delta)$ of the decomposition, we construct sets $F_A(x)$, $F_C(x)$ that serve as "funnels" to connect the neighborhood $N_B(x,\eta_2)$ to $M_5(x)$. In $N_B(x,\eta_2)$, the points corresponding to $(w_1,w_2) = (0,\eta_2)$ and $(0,-\eta_2)$ lie on the unstable manifold of the saddle $B(x)$, and the corresponding (exact) forward trajectories of (3.9) from these points approach the stable critical points $A(x)$ and $C(x)$ as $t \to \infty$, respectively (upon a suitable choice of the orientation of the eigenvector r_2 in §7).

To construct $F_A(x)$, let $z(x,t)$ denote the exact solution of $z_t = G(x,z)$, so that $z(x,0)$ corresponds to $(w_1,w_2) = (0,\eta_2)$. Because solutions of (3.9) depend continuously on x and initial data, and the fact that $N_A(x,1/K_a^2)$ is positively invariant, there exist constants $\eta > 0$, $T > 0$ (depending on h but not on x) such that

$$\{u \in R^2 \mid |u - z(x,T)| \le \eta\} \subset N_A\left(x, \frac{1}{K_a^2}\right)$$

for x satisfying (9.1).

Let L be an upper bound for the Lipschitz constant of G in M_0, and let $\kappa > 2L$. Suppose $u = u(x,t)$ is a δ-approximate solution of (3.9), and so satisfies (7.9). Then

$$\begin{aligned} \frac{d}{dt}|u-z|^2 e^{-2\kappa t} &= 2e^{-2\kappa t}(u-z)^T(-\kappa(u-z) + G(x,u) - G(x,z) + f) \\ &\le 2e^{-2\kappa t}((-\kappa + L)|u-z|^2 + |u-z|\delta) < 0, \end{aligned} \tag{9.4}$$

provided $2\delta < \kappa|u-z|$. From this we deduce the following.

LEMMA 9.2. *For x satisfying (9.1), define*

$$F_A(x) = \{u \in R^2 \mid |u - z(x,t)| \le \eta e^{\kappa(t-T)} \textit{ for some } t \in [0,T]\}.$$

Then, if $\delta < \kappa\eta e^{-\kappa T}/2$ and $\delta < 1/K_a^2$, the set $F_A(x) \cup N_A(x,1/K_a^2)$ is positively invariant for δ-approximate solutions of (3.9). Moreover, any δ- approximate solution lying in $F_A(x)$ must enter $N_A(x,1/K_a^2)$.

In a similar manner, we construct the set $F_C(x)$, and we may say that for δ satisfying the conditions of Lemma 9.2, the set

$$M_4(x) = F_A(x) \cup F_C(x) \cup M_5(x)$$

is positively invariant for δ-approximate solutions, and any δ-approximate solution lying in $M_4(x)$ must enter $M_5(x)$.

The remaining sets $M_1(x,\delta)$ through $M_3(x,\delta)$ will be defined by using the coordinates (w_1,w_2) in the neighborhood of the saddle point B. Let $K>0$ be a uniform bound for the Lipschitz constant of the change of variables from (w_1,w_2) to u in $N_B(x,\eta_2)$. Define

$$N_1(x)=\left\{u\in N_B(x,\eta_2)\mid |w_1|\le\frac{\eta e^{-\kappa T}}{K}\right\}.$$

By Lemma 9.1, if $\delta<\eta e^{-\kappa T}/KK_b$, then any δ-approximate solution of (3.9) which leaves $N_1(x)$ must do so at a point where $w_2=\eta_2$ or $-\eta_2$ and $|w_1|<\eta e^{-\kappa T}/K$. Thus this solution must enter one of the funnels in $M_4(x)$. We define

$$M_1(x)=N_1(x)\cup M_4(x).$$

Next, with reference to Lemma 9.1 we define

$$N_2(x,\delta)=\{u\in N_1(x)\mid |w_1|\le K_b\delta\},$$
$$N_3(x,\delta)=\{u\in N_1(x)\mid |w|_\infty\le K_b\delta\},$$

and put

$$M_2(x,\delta)=M_4(x)\cup N_2(x,\delta),\qquad M_3(x,\delta)=M_2(x,\delta)\setminus N_3(x,\delta).$$

For δ sufficiently small (independent of x), the sets $M_1(x)$, $M_2(x,\delta)$, and $M_3(x,\delta)$ are positively invariant for δ-approximate solutions. Note that we have

$$\begin{aligned}M_1(x)\setminus M_2(x,\delta)&\subset\{u\in N_1(x)\mid |w_1|\ge K_b\delta\},\\ M_2(x,\delta)\setminus M_3(x,\delta)&\subset\{u\in N_1(x)\mid |w|_\infty\le K_b\delta\}=N_3(x,\delta),\\ M_3(x,\delta)\setminus M_4(x)&\subset\{u\in N_1(x)\mid |w_1|\le K_b\delta\ \text{ and }\ |w_2|\ge K_b\delta\}.\end{aligned}$$

By Lemma 9.1, for δ sufficiently small, if $u(x,t)$ is any δ-approximate solution lying in $M_1(x)\setminus M_2(x,\delta)$, then $|w_1|$ is strictly decreasing, so u must enter $M_2(x,\delta)$. If u lies in $M_3(x,\delta)\setminus M_4(x)$, then $|w_2|$ is strictly increasing, so u must enter $M_4(x)$.

PROPOSITION 9.3. *Let $h>0$. Then there exist constants $\delta_0>0$, $K>0$ such that, if $\delta\le\delta_0$ and x satisfies* (9.1), *then:*

- *The sets $M_j(x,\delta)$ are positively invariant for δ-approximate solutions of* (3.9) *for $j=0,\dots,6$.*
- *For $j=0,1,3,4,$ and 5, every δ-approximate solution that lies in $M_j(x,\delta)$ must enter $M_{j+1}(x,\delta)$.*

Hence every recurrent point of a δ-approximate solution must lie either in $N_3(x,\delta)\supset M_2(x,\delta)\setminus M_3(x,\delta)$ or in $M_6(x,\delta)$, and there we have

$$\sup_{u\in N_3(x,\delta)}|G(x,u)|\le K\delta,\qquad \sup_{u\in M_6(x,\delta)}|G(x,u)|\le K\delta. \tag{9.5}$$

Proof. The inequality (9.5) follows easily from the definitions of $N_3(x,\delta)$ and $M_6(x,\delta)$ and Lemma 7.2. The only assertion remaining to be proved is, that for δ sufficiently small (depending on h but not on x), if $u(x,t)$ is a δ-approximate solution lying in $M_0(x)$, then u enters $M_1(x)$. But this is proved exactly the same as in the proof of Proposition 8.2, by using Proposition 3.7, this time taking

$$\mathcal{N}=\{(x,u)\in[-1,0]\times R^2\mid u\in M_1(x)\text{ if }x\text{ satisfies (9.1)}\}.$$

10. A semi-Morse decomposition: case (iii). In this section we study the behavior of δ-approximate solutions of (3.9) for x near x_m and x_M, the points at which saddle-node bifurcations occur in (3.9). In fact, it suffices to consider x near x_m (where the critical points B and C coalesce); the situation for x near x_M (where A and B coalesce) is entirely analogous. The constructions in this section are similar to those in §9. The semi-Morse decomposition can consist of as many as nine sets, $M_0(x,\delta), \ldots, M_8(x,\delta)$. A sketch appears in Fig. 8, with details of the neighborhood of the saddle-node pair in Figs. 9 and 10, corresponding to a fixed δ, for two different values of x near x_m with $x < x_m$. (Two other possibilities occur with $x > x_m$, when the critical points B and C do not exist.) As before, the numeral $j = 0, \ldots, 8$ in the figures indicate the region $M_j \setminus M_{j+1}$.

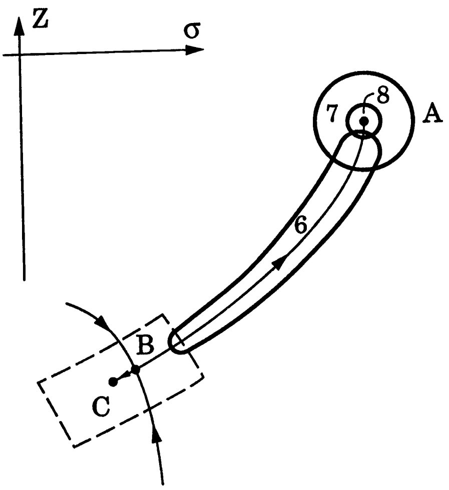


FIG. 8. *Semi-Morse decomposition, case* (iii).

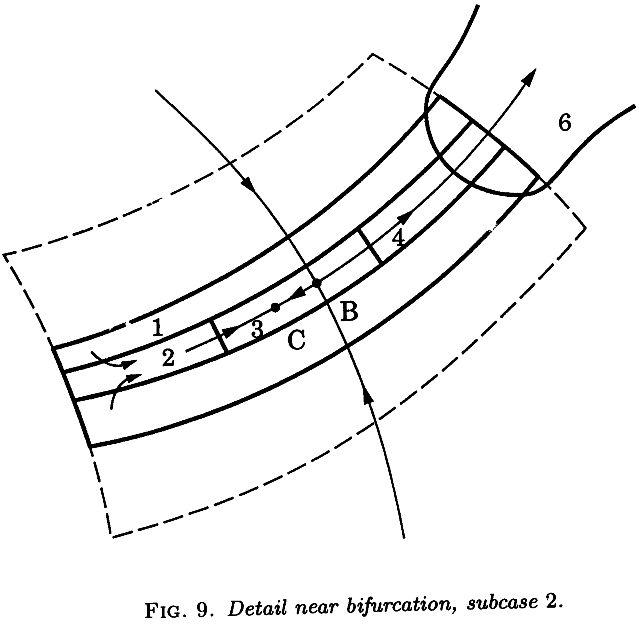


FIG. 9. *Detail near bifurcation, subcase* 2.

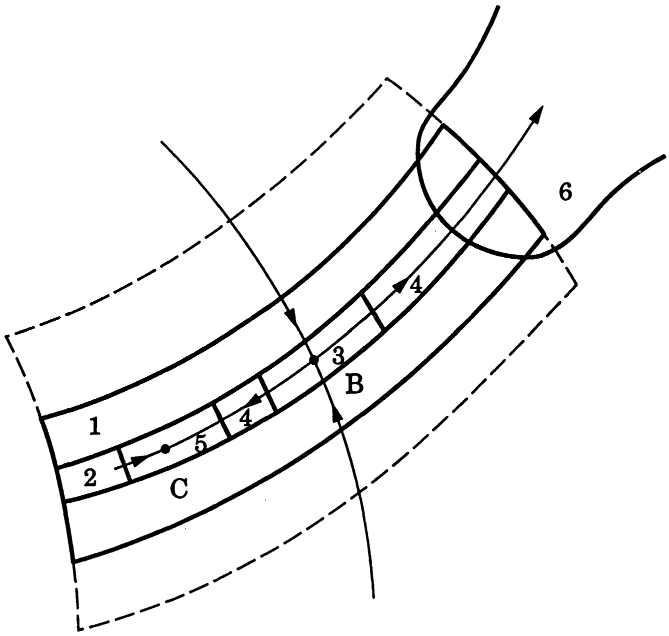


FIG. 10. *Detail near bifurcation, subcase* 3.

To begin, we invoke all the analysis of §7, case (c). In terms of the variables $W = (w_1, w_2, y)$ (recall $y = x - x_m$), the results in (7.33)–(7.37) have been obtained. We may summarize these as follows.

LEMMA 10.1. *There exist positive constants h_2, K_c, and h such that for $K_c\delta < h_2$ and $|x - x_m| < h$, if $u = u(x,t)$ is a δ-approximate solution of* (3.9) *satisfying $|w|_\infty = \max(|w_1|, |w_2|) \le h_2$, then*

$$\begin{array}{ll} |w_1| \ge K_c\delta & \textit{implies } |w_1| \textit{ is strictly decreasing;} \\ |w_2^2 + ay| \ge \max(K_c\delta, |w_1|) & \textit{implies } \operatorname{sgn}(w_{2\,t}) = \operatorname{sgn}(w_2^2 + ay). \end{array}$$

Moreover, if h is sufficiently small (*independent of x*) *we have*

$$\begin{aligned} &\frac{\partial}{\partial y}(w_2^2 + ay) \ge \frac{1}{2}a_2 > 0, \qquad \frac{\partial^2}{\partial w_2^2}(w_2^2 + ay) \ge 1 \quad \textit{for } |y| < h, \ |w_2| \le h_2, \\ &h_2^2 + a(h_2, y)y > 0 \quad \textit{for } |y| < h. \end{aligned} \tag{10.1}$$

The result of this lemma fixes the value of h to be used below and in previous sections. Now, for $|x - x_m| < h$, the smallest sets in the semi-Morse decomposition will be neighborhoods of the stable critical point $A(x)$. Using Lemma 8.1 we put

$$M_8(x,\delta) = N_A(x,\delta), \qquad M_7(x) = N_A\left(x, \frac{1}{K_a^2}\right).$$

From (7.10), (7.11), for $\delta < 1/K_a^2$, any δ-approximate solution lying in $M_7(x)$ must enter $M_8(x,\delta)$.

For the next step, we introduce a funnel $F_A(x)$, constructed in a manner entirely analogous to that in section 9, having the form in Lemma 9.2, where now $z(x,0)$ corresponds to the point $(w_1, w_2) = (0, h_2)$, whose (exact) forward trajectory under (3.9) approaches $A(x)$ as $t \to \infty$. We define

$$M_6(x) = F_A(x) \cup M_7(x).$$

Provided $\delta < \min(\kappa\eta e^{-\kappa T}/2, 1/K_a^2)$, the set $M_6(x)$ is positively invariant for δ-approximate solutions, and any such solution lying in $M_6(x)$ must enter $M_7(x)$.

The remaining constructions are carried out in the local coordinates (w_1, w_2, y) defined near the bifurcation point. Let $K > 0$ be a uniform bound for the Lipschitz constant of the change of variables from (u, x) to (w, y), and define, for $|x - x_m| < h$,

$$N_1(x) = \left\{ u \mid |w|_\infty \leq h_2 \text{ and } |w_1| \leq \frac{\eta e^{-\kappa T}}{K} \right\}.$$

By Lemma 10.1, if $K_c\delta < \eta e^{-\kappa T}/K$, then any δ-approximate solution which leaves $N_1(x)$ must do so at a point where $w_2 = h_2$; thus this solution enters $M_6(x)$. We define

$$M_1(x) = N_1(x) \cup M_6(x).$$

Then $M_1(x)$ is positively invariant for δ-approximate solutions, if δ is sufficiently small (independent of x). Moreover, by using Proposition 3.7 exactly as in the proof of Propositions 8.2 and 9.3, it is clear that for δ sufficiently small and $|x - x_m| < h$, any δ-approximate solution in $M_0(x)$ must enter $M_1(x)$.

Next, we define, for $|x - x_m| < h$,

$$N_2(x, \delta) = \{u \in N_1(x) \mid |w_1| \leq K_c\delta\},$$
$$M_2(x, \delta) = N_2(x, \delta) \cup M_6(x).$$

By Lemma 10.1, it is clear that $M_2(x, \delta)$ is positively invariant for δ-approximate solutions, and any such solution lying in $M_1(x)$ must enter $M_2(x, \delta)$, since $|w_1|$ is strictly decreasing in $M_1(x) \setminus M_2(x, \delta)$.

Now, for a δ-approximate solution lying in $N_2(x, \delta)$, Lemma 10.1 implies

$$\operatorname{sgn}(w_{2t}) = \operatorname{sgn}(w_2^2 + ay) \quad \text{provided } |w_2^2 + ay| \geq K_c\delta. \tag{10.2}$$

The remaining constructions subdivide into three cases (illustrated in Fig. 11).

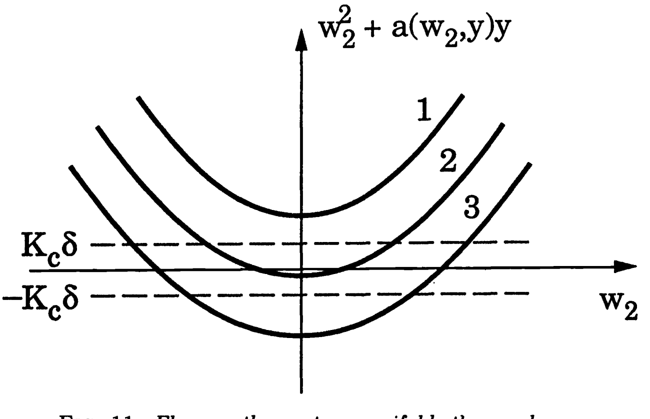


FIG. 11. *Flow on the center manifold: three subcases.*

(1) In the simplest case, $x = x_m + y$ is such that

$$w_2^2 + ay > K_c\delta \quad \text{for all } |w_2| \leq h_2.$$

In this case, the critical points B and C are absent, and w_2 is strictly increasing in $N_2(x,\delta)$. Thus every δ-approximate solution lying in $M_2(x,\delta)$ must enter $M_6(x)$. We define

$$M_j(x,\delta) = M_6(x) \quad \text{for } j = 3,\ 4, \text{ and } 5.$$

(2) In the next case (which includes $x = x_m$) we have

$$\{w_2 \mid |w_2^2 + ay| \leq K_c\delta\} = [\gamma_1, \gamma_2]$$

for some interval $[\gamma_1, \gamma_2]$ depending on x and δ. In this case, the critical points may or may not be absent. We define

$$\begin{aligned} M_3(x,\delta) &= M_6(x) \cup \{u \in N_2(x,\delta) \mid w_2 \geq \gamma_1\}, \\ M_4(x,\delta) &= M_6(x) \cup \{u \in N_2(x,\delta) \mid w_2 \geq \gamma_2\}, \\ M_5(x,\delta) &= M_6(x). \end{aligned}$$

It is clear that for δ sufficiently small, each $M_j(x,\delta)$ is positively invariant for δ-approximate solutions and every δ-approximate solution lying in $M_j(x,\delta)$ must enter $M_{j+1}(x,\delta)$, for $j = 2$, 4, and 5. Moreover, for $u \in M_3(x,\delta) \setminus M_4(x,\delta)$ we have $\gamma_1 \leq w_2 \leq \gamma_2$, so for some constant K we have

$$|G(x,u)| \leq K\delta \quad \text{for } u \in M_3(x,\delta) \setminus M_4(x,\delta).$$

(3) In the last case, we have

$$\{w_2 \mid |w_2^2 + ay| \leq K_c\delta\} = [\gamma_1, \gamma_2] \cup [\gamma_3, \gamma_4]$$

for some $\gamma_1 \leq \gamma_2 \leq \gamma_3 \leq \gamma_4$ depending on x and δ. For a δ-approximate solution lying in $N_2(x,\delta)$, w_2 is strictly increasing if $w_2 \leq \gamma_1$ or $w_2 \geq \gamma_4$, and w_2 is strictly decreasing if $\gamma_2 \leq w_2 \leq \gamma_3$. We put

$$\begin{aligned} M_3(x,\delta) &= M_6(x) \cup \{u \in N_2(x,\delta) \mid \gamma_1 \leq w_2\}, \\ M_4(x,\delta) &= M_6(x) \cup \{u \in N_2(x,\delta) \mid \gamma_1 \leq w_2 \leq \gamma_3 \text{ or } \gamma_4 \leq w_2\}, \\ M_5(x,\delta) &= M_6(x) \cup \{u \in N_2(x,\delta) \mid \gamma_1 \leq w_2 \leq \gamma_2\}. \end{aligned}$$

It is clear that for δ sufficiently small, each $M_j(x,\delta)$ is positively invariant for δ-approximate solutions, and every δ-approximate solution lying in $M_j(x,\delta)$ must enter $M_{j+1}(x,\delta)$ for $j = 2$ and 4. Moreover, for some constant K we clearly have

$$|G(x,u)| \leq K\delta \quad \text{for } u \in M_j(x,\delta) \setminus M_{j+1}(x,\delta) \quad \text{for } j = 3 \quad \text{and} \quad 5.$$

This finishes our analysis of case (iii), and concludes the proof of Proposition 3.11.

REFERENCES

[1] F. Brauer and J. A. Nohel, *The Qualitative Theory of Ordinary Differential Equations: An Introduction*, Benjamin, New York, 1969.

[2] C. C. Conley, *Isolated Invariant Sets and the Morse Index*, CBMS Regional Conference Series in Mathematics, No. 38, American Mathematical Society, 1978.

[3] R. Franzosa, *Index filtrations and the homology index braid for partially ordered Morse decompositions*, Trans. Amer. Math. Soc., 298 (1986), pp. 193–213.

[4] C. Guillopé and J.-C. Saut, *Existence and nonlinear stability of shearing motions of non-Newtonian fluids of Oldroyd type*, Math. Mod. Numer. Anal., 24 (1990), pp. 369–401.

[5] ———, *Existence results for the flow of viscoelastic fluids with a differential constitutive law*, Nonlinear Anal. TMA, 15 (1990), pp. 849–869.

[6] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math., 840, Springer-Verlag, New York, 1981.

[7] A. Kelly, *The stable, center-stable, center, center-unstable, and unstable manifolds*, in Transversal Mappings and Flows, R. Abraham and J. Robbin, eds., Benjamin, New York, 1967.

[8] D. S. Malkus, J. A. Nohel, and B. J. Plohr, *Analysis of new phenomena in shear flow of non-Newtonian fluids*, SIAM J. Appl. Math., 51 (1991), pp. 899–929.

[9] J. A. Nohel, R. L. Pego, and A. E. Tzavaras, *Stability of discontinuous steady states in shearing motions of a non-Newtonian fluid*, Proc. Roy. Soc. Edinburgh, 115A (1990), pp. 39–59.

[10] D. E. Norton, *A metric approach to the Conley decompostion theorem*, Villanova University, Villanova, PA, preprint, 1990.

SIAM J. MATH. ANAL.
Vol. 24, No. 4, pp. 943–967, July 1993

ANTIPLANE SHEARING MOTIONS OF A VISCO-PLASTIC SOLID*

J. M. GREENBERG[†] AND ANNE NOURI[‡]

Abstract. The authors consider antiplane shearing motions of an incompressible isotropic visco-plastic solid. The flow rule employed is a properly invariant generalization of Coulomb sliding friction and assumes a constant yield stress or threshold above which plastic flow occurs. In this model stresses above yield are possible; but when this condition obtains, the plastic flow rule forces the plastic strain to change so as to lower the stress levels in the material and dissipate energy. On the yield surface, the flow rule looks like the classical one for a rate independent elastic-perfectly plastic material when the velocity gradients are small enough but differs from the classical model for large gradients.

Key words. plastic waves, visco-plasticity, time-dependent problems

AMS subject classifications. 73E70, 73E60, 73E50

1. Introduction.

In this note we consider antiplane shearing motions of an incompressible isotropic visco-plastic solid. This work generalizes and compliments earlier work of Greenberg [1], [2], where he considered simple shearing flows for such materials. The flow rule we employ is a properly invariant generalization of Coulomb sliding friction and assumes a constant yield stress or threshold above which plastic flow occurs. As with most such theories, we assume a multiplicative decomposition of the deformation gradient into an elastic and plastic part, and we assume further that the deviatoric part of the Cauchy Stress tensor depends only on the elastic portion of the deformation gradient. For antiplane shearing motions this decomposition presents no precedence problems; i.e., does the elastic deformation precede the plastic or vice versa? One key feature of this model is that stresses above yield are possible. When this condition obtains, the plastic flow rule forces the plastic strain to change so as to lower the stress levels in the material and dissipate energy. The principal difficulty in formulating this model occurs when the stress is at yield. Motivated by results of Seidman [3], Utkin [4], and Filippov [5] on sliding modes induced by discontinuous vector fields, we are led to the flow rule advanced in (2.38). On the yield surface, this flow rule looks like the classical one for a rate independent elastic-perfectly plastic material when the velocity gradients are small enough but differs from the classical model for large gradients. This rule differentiates between loading and unloading and generates an energy identity which guarantees that uniqueness obtains for initial and initial-boundary value problems.

The organization of this paper is as follows. In §2 we develop the appropriate equations describing antiplane shearing flows in visco-plastic solids. Section 3 focuses on the uniqueness issue. Our basic estimate is that the energy associated with the difference between two solutions generated by the same data is nonincreasing. This estimate relies in an essential way on the definition of the plastic flow rule. In §4 we examine a one-dimensional signalling problem and discuss (1) the structure of this

* Received by the editors March 17, 1992; accepted for publication (in revised form) September 2, 1992.

[†] Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, Maryland 21228. This author's research was partially supported by the U.S. National Science Foundation and the U.S. Department of Energy.

[‡] Université de Nice, Parc Valrose, 06034 Nice Cedex, France. This author's research was partially supported by Centre National de la Recherche Scientifique, France.

solution, and (2) a procedure to analytically obtain an approximate solution. We also compare this solution with what obtains for the more studied model of a rate independent elastic-perfectly plastic material where uniqueness fails. Section 5 deals with a numerical experiment for a two-dimensional signalling problem in the corner domain $r > 0$ and $\pi/2 < \theta < 2\pi$. Here the stresses are singular as one approaches the corner and care must be taken in the implementation of the boundary conditions.

We note that in the last several years there have been a number of other efforts aimed at capturing the essence of plastic flows. Antman and Szymczak [6], [7] have advanced a finite deformation theory of such materials which is similar in spirit to ours but differs in a number of essential ways. Their model is formally rate independent where ours is not but their model also requires a history dependent strain hardening mechanism. The predictions of the two theories are often qualitatively different; these differences arise since in their model the imposition of large loads tends to elevate the yield stress and create a temporally constant permanent plastic deformation, whereas in our model such loading would generate a constant plastic deformation rate and thus a plastic deformation which varies linearly in time. This may be seen by examining the solution constructed in §4. Other efforts on elasto-plastic modelling may be found in Coleman and Owen [8], Buhite and Owen [9], Coleman and Hodgdon [10], and Owen [11].

2. Model development.

We say that a body is undergoing antiplane shear if material points $\xi = \xi_1\mathbf{e}_1 + \xi_2\mathbf{e}_2 + \xi_3\mathbf{e}_3$ move to $\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + x_3\mathbf{e}_3$ with

$$x_1 = \xi_1, \quad x_2 = \xi_2, \quad \text{and } x_3 = \xi_3 + \phi(\xi_1, \xi_2, t) \tag{2.1}$$

under the action of a Cauchy stress tensor of the form

$$\begin{aligned} T = &-\pi(\mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_2 + \mathbf{e}_3 \otimes \mathbf{e}_3) \\ &+(S_{11}\mathbf{e}_1 \otimes \mathbf{e}_1 + S_{22}\mathbf{e}_2 \otimes \mathbf{e}_2 + S_{33}\mathbf{e}_3 \otimes \mathbf{e}_3) \\ &+S_{31}(\mathbf{e}_1 \otimes \mathbf{e}_3 + \mathbf{e}_3 \otimes \mathbf{e}_1) + S_{32}(\mathbf{e}_2 \otimes \mathbf{e}_3 + \mathbf{e}_3 \otimes \mathbf{e}_2). \end{aligned} \tag{2.2)^1 $(}$$

Here, π is the hydrostatic pressure and S is the deviatoric stress tensor and satisfies

$$\text{trace}(S) = S_{11} + S_{22} + S_{33} = 0. \tag{2.3}$$

Relative to the above basis, the matrix representation of the Cauchy stress is given by

$$\mathcal{J} = -\pi \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} S_{11} & 0 & S_{31} \\ 0 & S_{22} & S_{32} \\ S_{31} & S_{32} & S_{33} \end{pmatrix}, \tag{2.4}$$

and relative to the same basis the deformation gradient is given by

$$\mathcal{F} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ F_{31} & F_{32} & 1 \end{pmatrix}, \tag{2.5}$$

[1] $\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, $\mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, and $\mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ are the standard basis elements for R^3 and $\mathbf{e}_i \otimes \mathbf{e}_j = \mathbf{e}_i\mathbf{e}_j{}^{\top}$ are the standard basis elements for linear operators from R^3 to R^3.

where

$$F_{31} = \frac{\partial \phi}{\partial x_1} \quad \text{and} \quad F_{32} = \frac{\partial \phi}{\partial x_2}. \tag{2.6}$$

Noting that matrices

$$\mathcal{F}_{(a,b)} \overset{\text{def}}{:=} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ a & b & 1 \end{pmatrix}$$

satisfy the commutation relation

$$\mathcal{F}_{(a_1,b_1)}\mathcal{F}_{(a_2,b_2)} = \mathcal{F}_{(a_2,b_2)}\mathcal{F}_{(a_1,b_1)} = \mathcal{F}_{(a_1+a_2,b_1+b_2)}, \tag{2.7}$$

we feel justified in decomposing the deformation gradient $\mathcal{F}$ into its elastic and plastic parts $\mathcal{E}$ and $\mathcal{P}$ by

$$\mathcal{E} \overset{\text{def}}{:=} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ e_{31} & e_{32} & 1 \end{pmatrix} \quad \text{and} \quad \mathcal{P} \overset{\text{def}}{:=} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ p_{31} & p_{32} & 1 \end{pmatrix}, \tag{2.8}$$

where

$$\mathcal{F} = \mathcal{E}\mathcal{P} = \mathcal{P}\mathcal{E} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ e_{31}+p_{31} & e_{32}+p_{32} & 1 \end{pmatrix}. \tag{2.9}$$

For such antiplane shear flows one need not make any assumption about the precedence of the elastic and plastic parts of the flow.

Our basic constitutive assumption is that under a change of reference frame E transforms in the same way as F and that the deviatoric stress S is an isotropic, frame indifferent, trace free function of the elastic deformation gradient E.[2] The constraint that S is an isotropic, frame indifferent function of E implies that $\mathcal{S}$ must have the functional form

$$\mathcal{S} = \alpha I + \beta \mathcal{E}\mathcal{E}^{\top} + \gamma \mathcal{E}^{-\top}\mathcal{E}^{-1} \tag{2.10}$$

or

(2.11)[3]

$$\mathcal{S} = \alpha \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \beta \begin{pmatrix} 1 & 0 & e_{31} \\ 0 & 1 & e_{32} \\ e_{31} & e_{32} & 1+e_{31}^2+e_{32}^2 \end{pmatrix} + \gamma \begin{pmatrix} 1+e_{31}^2 & e_{31}e_{32} & -e_{31} \\ e_{31}e_{32} & 1+e_{32}^2 & -e_{32} \\ -e_{31} & -e_{32} & 1 \end{pmatrix},$$

where α, β, and γ are functions of the invariants of $\mathcal{E}\mathcal{E}^{\top}$, in this case the scalar $e_{31}^2 + e_{32}^2$. Equation (2.4) implies that $\mathcal{S}_{21} = \mathcal{S}_{12} = 0$ and this, in turn, implies that $\gamma \equiv 0$ while the condition that trace$\mathcal{S} = 0$ implies that $\alpha = -\beta(1 + ((e_{31}^2 + e_{32}^2)/3))$. Combining these identities with (2.11) yields

$$\mathcal{S} = \beta \begin{pmatrix} -\frac{1}{3}(e_{31}^2 + e_{32}^2) & 0 & e_{31} \\ 0 & -\frac{1}{3}(e_{31}^2 + e_{32}^2) & e_{32} \\ e_{31} & e_{32} & \frac{2}{3}(e_{31}^2 + e_{32}^2) \end{pmatrix}. \tag{2.12}$$

[2] E is the tensor whose matrix representation relative to the basis elements $\mathbf{e_i} \otimes \mathbf{e_j}$ is given by $(2.8)_1$.

[3] For details see Gurtin [12].

In the sequel we shall assume that β is a positive constant. Equation (2.12) implies that we may regard the elements S_{31} and S_{32} as basic descriptors of our system. In terms of these $\mathcal{S}$ and $\mathcal{E}$ take the form

$$\mathcal{S} = \begin{pmatrix} -\frac{1}{3\beta}(S_{31}^2 + S_{32}^2) & 0 & S_{31} \\ 0 & -\frac{1}{3\beta}(S_{31}^2 + S_{32}^2) & S_{32} \\ S_{31} & S_{32} & \frac{2}{3\beta}(S_{31}^2 + S_{32}^2) \end{pmatrix} \tag{2.13}$$

and

$$\mathcal{E} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{S_{31}}{\beta} & \frac{S_{32}}{\beta} & 1 \end{pmatrix}. \tag{2.14}$$

We now turn to the equations of motion. Equation (2.1) implies that the Eulerian velocity field $\mathbf{u}$ is of the form

$$\mathbf{u} = u(x_1, x_2, t_1)\mathbf{e_3}, \tag{2.15}$$

where

$$u(x_1, x_2, t_1) = \frac{\partial \phi}{\partial t_1}(x_1, x_2, t_1); \tag{2.16}$$

and (2.16), when combined with (2.6), implies that

$$\frac{\partial F_{31}}{\partial t_1} - \frac{\partial u}{\partial x_1} = 0 \tag{2.17}$$

and

$$\frac{\partial F_{32}}{\partial t_1} - \frac{\partial u}{\partial x_2} = 0. \tag{2.18}$$

Additionally, (2.9) and (2.14) imply that

$$F_{31} = \frac{S_{31}}{\beta} + p_{31} \tag{2.19}$$

and

$$F_{32} = \frac{S_{32}}{\beta} + p_{32}. \tag{2.20}$$

Balance of momentum in the $\mathbf{e_1}$ and $\mathbf{e_2}$ directions implies that

$$\frac{\partial}{\partial x_1}\left(\pi + \frac{(S_{31}^2 + S_{32}^2)}{3\beta}\right) = \frac{\partial}{\partial x_2}\left(\pi + \frac{(S_{31}^2 + S_{32}^2)}{3\beta}\right) = 0 \tag{2.21}$$

or equivalently that

$$\pi = \pi_0(x_3, t) - \frac{(S_{31}^2 + S_{32}^2)}{3\beta}(x_1, x_2, t), \tag{2.22}$$

whereas balance of momentum in the $\mathbf{e}_3$ direction yields

$$\rho_0 \frac{\partial u}{\partial t_1} - \frac{\partial S_{31}}{\partial x_1} - \frac{\partial S_{32}}{\partial x_2} = -\frac{\partial \pi_0}{\partial x_3}. \tag{2.23}$$

Here, ρ_0 is the constant mass density of the material. Since $\partial\pi_0/\partial x_3$ depends on x_3 and t_1, whereas all quantities on the left-hand side of (2.23) depend only on x_1, x_2, and t_1, we conclude that for antiplane shearing flows $\partial\pi_0/\partial x_3$ is independent of x_3. In what follows we shall assume this quantity is zero.

We now turn our attention to "yield condition" and the flow rule for the plastic strain tensor $\mathcal{P}$ of $(2.8)_2$. We assume that yield is determined by whether the scalar $S_{31}^2 + S_{32}^2$ exceeds a threshold S_y^2 or not. This assumption relies on the special form of $\mathcal{S}$ (see (2.13)) and is equivalent to a yield criteria determined by the norm of $\mathcal{S}$, where

$$\|\mathcal{S}\|^2 \overset{\text{def}}{:=} \mathcal{S}_{ij}\mathcal{S}_{ij} = 2(S_{31}^2 + S_{32}^2) + \frac{2}{3\beta^2}(S_{31}^2 + S_{32}^2)^2 \tag{2.24}$$

or one based on the maximum shear stress

$$S_*^2 \overset{\text{def}}{:=} \max_{\{\mathbf{e}|\mathbf{e}\cdot\mathbf{e}=1\}} \|\mathcal{S}\mathbf{e} - (\mathcal{S}\mathbf{e}\cdot\mathbf{e})\mathbf{e}\|^2. \tag{2.25}$$

In the sequel we let H denote the Heaviside function

$$H(x) \overset{\text{def}}{:=} \begin{cases} 0, & x < 0 \\ 1, & x > 0 \end{cases} \tag{2.26}$$

and define ψ_1 and ψ_2 by

$$\psi_1 = \frac{1}{2}\int_{-\infty}^{(S_{31}^2+S_{32}^2)} H(x - S_y^2)\, dx \tag{2.27}$$

and

$$\psi_2 = \int_{-\infty}^{\sqrt{S_{31}^2+S_{32}^2}} H(x - S_y)\, dx, \tag{2.28}$$

where $S_y > 0$ is the "yield stress."

We shall confine our attention to the Coulomb type sliding law

$$\frac{\partial p_{31}}{\partial t_1} = \frac{1}{\beta T_0}\frac{\partial \psi_1}{\partial S_{31}} = \frac{S_{31}}{\beta T_0} H(S_{31}^2 + S_{32}^2 - S_y^2) \tag{2.29}$$

and

$$\frac{\partial p_{32}}{\partial t_1} = \frac{1}{\beta T_0}\frac{\partial \psi_1}{\partial S_{32}} = \frac{S_{32}}{\beta T_0} H(S_{31}^2 + S_{32}^2 - S_y^2), \tag{2.30}$$

though much of what we say applies equally well to the flow rule

$$\frac{\partial p_{31}}{\partial t_1} = \frac{S_y}{\beta T_0}\frac{\partial \psi_2}{\partial S_{31}} = \frac{S_y S_{31}}{\beta T_0\sqrt{S_{31}^2 + S_{32}^2}} H\left(\sqrt{S_{31}^2 + S_{32}^2} - S_y\right) \tag{2.31}$$

and

$$\frac{\partial p_{32}}{\partial t_1} = \frac{S_y}{\beta T_0}\frac{\partial \psi_2}{\partial S_{32}} = \frac{S_y S_{32}}{\beta T_0\sqrt{S_{31}^2 + S_{32}^2}} H\left(\sqrt{S_{31}^2 + S_{32}^2} - S_y\right). \tag{2.32}$$

The constant β is the shear modulus in (2.12), S_y is the yield stress, and $T_0 > 0$ is a fixed relaxation time. The flow rule is defined for $S_{31}^2 + S_{32}^2 \neq S_y^2$ and the problem remains to define it on the yield surface.

We first note that if $S_{31}^2 + S_{32}^2 \neq S_y^2$, we can combine (2.17)–(2.20) and (2.29) and (2.30) to obtain the following system for S_{31}, S_{32}, and u:

$$\frac{1}{\beta}\frac{\partial S_{31}}{\partial t_1} - \frac{\partial u}{\partial x_1} = \frac{-S_{31}H(S_{31}^2 + S_{32}^2 - S_y^2)}{\beta T_0}, \tag{2.33}$$

$$\frac{1}{\beta}\frac{\partial S_{32}}{\partial t_1} - \frac{\partial u}{\partial x_2} = \frac{-S_{32}H(S_{31}^2 + S_{32}^2 - S_y^2)}{\beta T_0}, \tag{2.34}$$

and

$$\rho_0\frac{\partial u}{\partial t_1} - \frac{\partial S_{31}}{\partial x_1} - \frac{\partial S_{32}}{\partial x_2} = 0. \tag{2.35}$$

Equations (2.33) and (2.34) imply that for $S_{31}^2 + S_{32}^2 \neq S_y^2$,

$$\begin{aligned}\frac{\partial}{\partial t_1}(S_{31}^2 + S_{32}^2) &= 2\beta\left(S_{31}\frac{\partial u}{\partial x_1} + S_{32}\frac{\partial u}{\partial x_2}\right)\\ &\quad -\frac{2}{T_0}(S_{31}^2 + S_{32}^2)H(S_{31}^2 + S_{32}^2 - S_y^2),\end{aligned} \tag{2.36}$$

and (2.36), together with the results of [3], [4], [5], motivates our extension of the flow rule on the yield surface $S_{31}^2 + S_{32}^2 = S_y^2$. We extend (2.29) and (2.30) to the yield surface $S_{31}^2 + S_{32}^2 = S_y^2$ by

$$\frac{\partial p_{31}}{\partial t_1} = \frac{\alpha S_{31}}{\beta T_0} \quad\text{and}\quad \frac{\partial p_{32}}{\partial t_1} = \frac{\alpha S_{32}}{\beta T_0}, \tag{2.37}$$

where

$$\alpha = \begin{cases} 1 \ \text{ if } S_{31}^2 + S_{32}^2 = S_y^2 \ \text{ and } \ S_{31}\dfrac{\partial u}{\partial x_1} + S_{32}\dfrac{\partial u}{\partial x_2} > \dfrac{S_y^2}{\beta T_0},^{4} \\ \beta T_0\left(S_{31}\dfrac{\partial u}{\partial x_1} + S_{32}\dfrac{\partial u}{\partial x_2}\right)\Big/ S_y^2 \ \text{ if } S_{31}^2 + S_{32}^2 = S_y^2 \ \text{ and} \\ 0 \le S_{31}\dfrac{\partial u}{\partial x_1} + S_{32}\dfrac{\partial u}{\partial x_2} \le \dfrac{S_y^2}{\beta T_0}, \\ 0 \ \text{ if } S_{31}^2 + S_{32}^2 = S_y^2 \ \text{ and } \ S_{31}\dfrac{\partial u}{\partial x_1} + S_{32}\dfrac{\partial u}{\partial x_2} < 0. \end{cases} \tag{2.38}$$

[4] The relations (2.37) and (2.38) transform in a frame indifferent fashion.

In the sequel we shall confine our attention to the extended flow rule (2.29), (2.30), (2.37) and (2.38). The relevant equations are

$$\frac{1}{\beta}\frac{\partial S_{31}}{\partial t_1} - \frac{\partial u}{\partial x_1} = -\frac{\alpha S_{31}}{\beta T_0}, \tag{2.39}$$

$$\frac{1}{\beta}\frac{\partial S_{32}}{\partial t_1} - \frac{\partial u}{\partial x_2} = -\frac{\alpha S_{32}}{\beta T_0}, \tag{2.40}$$

$$\rho_0\frac{\partial u}{\partial t_1} - \frac{\partial S_{31}}{\partial x_1} - \frac{\partial S_{32}}{\partial x_2} = 0, \tag{2.41}$$

where now

$$\alpha = \begin{cases} 1 \quad \text{if } S_{31}^2 + S_{32}^2 > S_y^2, \\ 1 \quad \text{if } S_{31}^2 + S_{32}^2 = S_y^2 \quad \text{and} \quad \dfrac{\beta T_0}{S_y^2}\left(S_{31}\dfrac{\partial u}{\partial x_1} + S_{32}\dfrac{\partial u}{\partial x_2}\right) > 1, \\ \dfrac{\beta T_0}{S_y^2}\left(S_{31}\dfrac{\partial u}{\partial x_1} + S_{32}\dfrac{\partial u}{\partial x_2}\right) \quad \text{if } S_{31}^2 + S_{32}^2 = S_y^2 \quad \text{and} \\ 0 \le \dfrac{\beta T_0}{S_y^2}\left(S_{31}\dfrac{\partial u}{\partial x_1} + S_{32}\dfrac{\partial u}{\partial x_2}\right) \le 1, \\ 0 \quad \text{if } S_{31}^2 + S_{32}^2 = S_y^2 \quad \text{and} \quad \dfrac{\beta T_0}{S_y^2}\left(S_{31}\dfrac{\partial u}{\partial x_1} + S_{32}\dfrac{\partial u}{\partial x_2}\right) < 0, \\ 0 \quad \text{if } S_{31}^2 + S_{32}^2 < S_y^2, \end{cases} \tag{2.42}$$

and these are solved together with appropriate initial and boundary conditions. Having solved the above system for S_{31}, S_{32}, and u we recover the deformation gradients F_{31} and F_{32} by solving

$$\frac{\partial F_{31}}{\partial t_1} - \frac{\partial u}{\partial x_1} = 0 \quad \text{and} \quad \frac{\partial F_{32}}{\partial t_1} - \frac{\partial u}{\partial x_2} = 0 \tag{2.43}$$

together with appropriate initial conditions. The plastic strains p_{31} and p_{32} are then given by

$$p_{31} = F_{31} - \frac{S_{31}}{\beta} \quad \text{and} \quad p_{31} = F_{32} - \frac{S_{32}}{\beta}. \tag{2.44}$$

These equations should be contrasted with what obtains in the more commonly studied theory of rate independent elastic-perfectly plastic materials. In that theory (2.37), (2.39)–(2.41), (2.43) and (2.44) still hold but α is given by

(2.45)

$$\alpha = \begin{cases} \frac{\beta T_0}{S_y^2}\left(S_{31}\frac{\partial u}{\partial x_1} + S_{32}\frac{\partial u}{\partial x_2}\right) & \text{if } S_{31}^2 + S_{32}^2 = S_y^2 \quad \text{and} \\ 0 \leq \frac{\beta T_0}{S_y^2}\left(S_{31}\frac{\partial u}{\partial x_1} + S_{32}\frac{\partial u}{\partial x_2}\right), & \\ 0 \quad \text{if } S_{31}^2 + S_{32}^2 = S_y^2 \quad \text{and} \quad \frac{\beta T_0}{S_y^2}\left(S_{31}\frac{\partial u}{\partial x_1} + S_{32}\frac{\partial u}{\partial x_2}\right) < 0, & \\ 0 \quad \text{if } S_{31}^2 + S_{32}^2 < S_y^2. & \end{cases}$$

The unboundedness of α on the yield surface $S_{31}^2 + S_{32}^2 = S_y^2$ presents difficulties not encountered in our model. In particular, across nonstationary shocks where F_{31}, F_{32}, u, S_{31}, and S_{32} experience jump discontinuities, we must admit jumps in the plastic strains p_{31} and p_{32}. The reason for this is that in the classical rate independent theory—α as in (2.45)—we must allow "dirac" type singularities in the terms $\alpha S_{31}/\beta T_0$ and $\alpha S_{32}/\beta T_0$ and therefore, we cannot conclude that

$$cn_1[p_{31}] = cn_2[p_{32}] = 0. \tag{2.46}$$

Here, c is the normal velocity of the shock wave and $\mathbf{n}=(n_1, n_2)$ is the unit normal to the shock. In our model α is bounded, no "dirac" type singularities arise in the terms $\alpha S_{31}/\beta T_0$ and $\alpha S_{32}/\beta T_0$, and thus (2.46) holds. This implies that with our model all nonstationary shocks satisfy $c^2 = 1$; that is, they propagate with the speed of elastic signals. With our model, the only surfaces across which the plastic strains can jump are stationary, i.e., $c = 0$. Such jumps are also allowed in the classical theory.

We conclude this section by writing down a dimensionless version (2.39)–(2.44). We let

(2.47)

$$x = \sqrt{\frac{\rho_0}{\beta}}\frac{x_1}{T_0}, \quad y = \sqrt{\frac{\rho_0}{\beta}}\frac{x_2}{T_0}, \quad t = \frac{t_1}{T_0}$$

$$v = \sqrt{\frac{\rho_0}{\beta}}u, \quad \tau_{31} = \frac{S_{31}}{\beta}, \quad \tau_{32} = \frac{S_{32}}{\beta}, \quad \text{and } \tau_y = \frac{S_y}{\beta}$$

and observe that (2.39)–(2.42) transform to

$$\frac{\partial \tau_{31}}{\partial t} - \frac{\partial v}{\partial x} = -\hat{\alpha}\tau_{31}, \tag{2.48}$$

$$\frac{\partial \tau_{32}}{\partial t} - \frac{\partial v}{\partial y} = -\hat{\alpha}\tau_{32}, \tag{2.49}$$

$$\frac{\partial v}{\partial t} - \frac{\partial \tau_{31}}{\partial x} - \frac{\partial \tau_{32}}{\partial y} = 0, \tag{2.50}$$

where

(2.51)
$$\hat{\alpha} = \begin{cases} 1 \quad \text{if } \tau_{31}^2 + \tau_{32}^2 > \tau_y^2, \\ 1 \quad \text{if } \tau_{31}^2 + \tau_{32}^2 = \tau_y^2 \quad \text{and} \quad \dfrac{1}{\tau_y^2}\left(\tau_{31}\dfrac{\partial v}{\partial x_1} + \tau_{32}\dfrac{\partial v}{\partial x_2}\right) > 1, \\ \dfrac{1}{\tau_y^2}\left(\tau_{31}\dfrac{\partial v}{\partial x_1} + \tau_{32}\dfrac{\partial v}{\partial x_2}\right) \quad \text{if } \tau_{31}^2 + \tau_{32}^2 = \tau_y^2 \text{ and} \\ 0 \leq \dfrac{1}{\tau_y^2}\left(\tau_{31}\dfrac{\partial v}{\partial x_1} + \tau_{32}\dfrac{\partial v}{\partial x_2}\right) \leq 1, \\ 0 \quad \text{if } \tau_{31}^2 + \tau_{32}^2 = \tau_y^2 \quad \text{and} \quad \dfrac{1}{\tau_y^2}\left(\tau_{31}\dfrac{\partial v}{\partial x_1} + \tau_{32}\dfrac{\partial v}{\partial x_2}\right) < 0, \\ 0 \quad \text{if } \tau_{31}^2 + \tau_{32}^2 < \tau_y^2. \end{cases}$$

The transformed versions of (2.43) and (2.44) are

(2.52)
$$\frac{\partial F_{31}}{\partial t} - \frac{\partial v}{\partial x} = 0 \quad \text{and} \quad \frac{\partial F_{32}}{\partial t} - \frac{\partial v}{\partial y} = 0$$

and

(2.53)
$$p_{31} = F_{31} - \tau_{31} \quad \text{and} \quad p_{32} = F_{32} - \tau_{32}.$$

3. Uniqueness results.

Our task in this section is to establish the following

THEOREM 3.1. *Let Ω be an open domain in R^2 with smooth boundary $\partial\Omega$. Then, there is at most one piecewise smooth,*[5] *$L^2_{\text{loc}}(\Omega)$ solution $(\tau_{31}, \tau_{32}, v)$ to* (2.48)–(2.51) *satisfying*

(3.1)
$$\lim_{t\to 0^+} (\tau_{31}, \tau_{32}, v)(x,y,t) = (\tau_{31}^0, \tau_{32}^0, v^0)(x,y),$$

(3.2)
$$\lim_{(x,y)\in\Omega;(x,y)\to\partial\Omega_1} (n_1\tau_{31} + n_2\tau_{32})(x,y,t) = f_1(x,y,t),$$

(3.3)
$$\lim_{(x,y)\in\Omega;(x,y)\to\partial\Omega_2} v(x,y,t) = f_2(x,y,t).$$

Here $\partial\Omega = \partial\Omega_1 \cup \partial\Omega_2$, $\partial\Omega_1 \cap \partial\Omega_2$ is at worst a finite collection of points, $\mathbf{n}=(n_1,n_2)$ is the unit exterior normal to $\partial\Omega_1$, and the f_i's are smooth functions in $L^2_{\text{loc}}(\partial\Omega_i \times [0,\infty))$.

Proof. We first note that if $(\tau_{31}^b, \tau_{32}^b, v^b)$ and $(\tau_{31}^a, \tau_{32}^a, v^a)$ are two solutions to (2.48)–(2.51), then their differences satisfy

(3.4)
$$\frac{\partial}{\partial t}(\tau_{31}^b - \tau_{31}^a) - \frac{\partial}{\partial x}(v^b - v^a) = -(\hat{\alpha}^b\tau_{31}^b - \hat{\alpha}^a\tau_{31}^a),$$

(3.5)
$$\frac{\partial}{\partial t}(\tau_{32}^b - \tau_{32}^a) - \frac{\partial}{\partial y}(v^b - v^a) = -(\hat{\alpha}^b\tau_{32}^b - \hat{\alpha}^a\tau_{32}^a),$$

and

(3.6)
$$\frac{\partial}{\partial t}(v^b - v^a) - \frac{\partial}{\partial x}(\tau_{31}^b - \tau_{31}^a) - \frac{\partial}{\partial y}(\tau_{32}^b - \tau_{32}^a) = 0.$$

[5] This formulation admits shocks which propagate with normal velocity c satisfying $c^2 = 1$.

Here, $\hat{\alpha}^b$ and $\hat{\alpha}^a$ represent the bounded function $\hat{\alpha}$ defined in (2.51) evaluated at $(\tau^b_{31}, \tau^b_{32}, v^b)$ and $(\tau^a_{31}, \tau^a_{32}, v^a)$, respectively. The last three identities imply that

$$\begin{aligned}
&\frac{1}{2}\frac{\partial}{\partial t}\left[(\tau^b_{31}-\tau^a_{31})^2+(\tau^b_{32}-\tau^a_{32})^2+(v^b-v^a)^2\right]\\
&\quad-\frac{\partial}{\partial x}\left[(\tau^b_{31}-\tau^a_{31})(v^b-v^a)\right]-\frac{\partial}{\partial y}\left[(\tau^b_{32}-\tau^a_{32})(v^b-v^a)\right]\\
&\quad=-\left[(\tau^b_{31}-\tau^a_{31})(\hat{\alpha}^b\tau^b_{31}-\hat{\alpha}^a\tau^a_{31})+(\tau^b_{32}-\tau^a_{32})(\hat{\alpha}^b\tau^b_{32}-\hat{\alpha}^a\tau^a_{32})\right].
\end{aligned}\tag{3.7}$$

We now claim that

$$p \overset{\text{def}}{:=} (\tau^b_{31}-\tau^a_{31})(\hat{\alpha}^b\tau^b_{31}-\hat{\alpha}^a\tau^a_{31})+(\tau^b_{32}-\tau^a_{32})(\hat{\alpha}^b\tau^b_{32}-\hat{\alpha}^a\tau^a_{32})\tag{3.8}$$

is nonnegative. In verifying this assertion there is no loss in generality in assuming that

$$0 \le \hat{\alpha}^a \le \hat{\alpha}^b \le 1.\tag{3.9}$$

We first note that p may be rewritten as

$$\begin{aligned}
p = {}&\hat{\alpha}^a\left[(\tau^b_{31}-\tau^a_{31})^2+(\tau^b_{32}-\tau^a_{32})^2\right]\\
&+(\hat{\alpha}^b-\hat{\alpha}^a)\left[(\tau^b_{31})^2-\tau^a_{31}\tau^b_{31}+(\tau^b_{32})^2-\tau^a_{32}\tau^b_{32}\right].
\end{aligned}\tag{3.10}$$

If $\hat{\alpha}^a = 0$, then $(\tau^a_{31})^2+(\tau^a_{32})^2 \le \tau_y^2$ and $\tau^a_{31}\tau^b_{31}+\tau^a_{32}\tau^b_{32} \le \tau_y\sqrt{(\tau^b_{31})^2+(\tau^b_{32})^2}$ and, therefore, (3.10) implies that

$$p \ge \hat{\alpha}^b\sqrt{(\tau^b_{31})^2+(\tau^b_{32})^2}\left(\sqrt{(\tau^b_{31})^2+(\tau^b_{32})^2}-\tau_y\right).\tag{3.11}$$

If $\hat{\alpha}^b = 0$, then (3.10) implies that $p = 0$, whereas if $0 < \hat{\alpha}^b \le 1$, (2.51) implies that $\sqrt{(\tau^b_{31})^2+(\tau^b_{32})^2} \ge \tau_y$, and (3.11) then yields $p \ge 0$. We now turn to the case where $0 < \hat{\alpha}^a \le \hat{\alpha}^b \le 1$. If $\hat{\alpha}^b = \hat{\alpha}^a$, the nonnegativity of p follows from (3.10), and thus to complete the verification that $p \ge 0$ it suffices to consider the case where $0 < \hat{\alpha}^a < \hat{\alpha}^b \le 1$. Here we know that $(\tau^a_{31})^2+(\tau^a_{32})^2 = \tau_y^2$ and $(\tau^b_{31})^2+(\tau^b_{32})^2 \ge \tau_y^2$. The former identity, along with (3.10) and $\tau^a_{31}\tau^b_{31}+\tau^a_{32}\tau^b_{32} \le \tau_y\sqrt{(\tau^b_{31})^2+(\tau^b_{32})^2}$, implies that

$$\begin{aligned}
p \ge {}&\hat{\alpha}^a\left[(\tau^b_{31}-\tau^a_{31})^2+(\tau^b_{32}-\tau^a_{32})^2\right]\\
&+(\hat{\alpha}^b-\hat{\alpha}^a)\sqrt{(\tau^b_{31})^2+(\tau^b_{32})^2}\left(\sqrt{(\tau^b_{31})^2+(\tau^b_{32})^2}-\tau_y\right),
\end{aligned}\tag{3.12}$$

and (3.12), $0 < \hat{\alpha}^a < \hat{\alpha}^b \le 1$, and $(\tau^b_{31})^2+(\tau^b_{32})^2 \ge \tau_y^2$ complete the proof of the assertion that p is nonnegative.

For any $(x_0, y_0) \in I\!R^2$, $r_0 > 0$, $T > 0$, and $0 \le t \le T$ we let

$$C(x_0, y_0, r_0, t) \overset{\text{def}}{:=} \{(x,y) | (x-x_0)^2+(y-y_0)^2 < (r_0+T-t)^2\}.\tag{3.13}$$

The identity (3.7) implies that if $(\tau^b_{31}, \tau^b_{32}, v^b)$ and $(\tau^a_{31}, \tau^a_{32}, v^a)$ are two solutions of (2.48)-(2.51) taking on the same data (3.1)–(3.3), then

$$\frac{1}{2}\int_{C(x_0,y_0,r_0,t)\cap\Omega}((\tau^b_{31}-\tau^a_{31})^2+(\tau^b_{32}-\tau^a_{32})^2+(v^b-v^a)^2)\,dx\,dy\tag{3.14}$$

$$
\begin{aligned}
&+\int_0^T \left(\int_{C(x_0,y_0,r_0,t)\cap\Omega} p(x,y,t)\,dx\,dy\right) dt\\
&+\int_0^T \left(\int_{\partial C(x_0,y_0,r_0,T)\cap\Omega} \left(\frac{1}{2}((\tau_{31}^b-\tau_{31}^a)^2+(\tau_{32}^b-\tau_{32}^a)^2+(v^b-v^a)^2)\right.\right.\\
&\left.\left.-(v^b-v^a)\left(\frac{(x-x_0)(\tau_{31}^b-\tau_{31}^a)}{(r_0+T-t)}+\frac{(y-y_0)(\tau_{32}^b-\tau_{32}^a)}{(r_0+T-t)}\right)\right)ds\right)dt\\
&=0.
\end{aligned}
$$

Here,

$$\partial C(x_0,y_0,r_0,t)=\{(x,y)|(x-x_0)^2+(y-y_0)^2=(r_0+T-t)^2\}. \tag{3.15}$$

The vector $((x-x_0)/(r_0+T-t),(y-y_0)/(r_0+T-t))$ is the unit exterior normal to $\partial C(x_0,y_0,r_0,t)$, and ds is arc length along $\partial C(x_0,y_0,r_0,t)$. Since

$$
\begin{aligned}
&-(v^b-v^a)\left(\frac{(x-x_0)(\tau_{31}^b-\tau_{31}^a)}{(r_0+T-t)}+\frac{(y-y_0)(\tau_{32}^b-\tau_{32}^a)}{(r_0+T-t)}\right)\\
&\geq -|v^b-v^a|\sqrt{(\tau_{31}^b-\tau_{31}^a)^2+(\tau_{32}^b-\tau_{32}^a)^2}\\
&\geq -\frac{1}{2}((\tau_{31}^b-\tau_{31}^a)^2+(\tau_{32}^b-\tau_{32}^a)^2+(v^b-v^a)^2)
\end{aligned}
\tag{3.16}
$$

and since $p\geq 0$, we see that all three integrals in (3.15) are nonnegative and their sum is zero. From this we obtain

$$\int_{C(x_0,y_0,r_0,T)\cap\Omega}((\tau_{31}^b-\tau_{31}^a)^2+(\tau_{32}^b-\tau_{32}^a)^2+(v^b-v^a)^2)dxdy=0, \tag{3.17}$$

which is the desired uniqueness result.

4. A signalling problem.

In this section we consider an elementary one-dimensional signalling problem for the normalized system (2.48)–(2.53). The solution is of the form

$$(\tau_{31},\tau_{32},v)=(\tau(x,t),0,v(x,t)),\qquad 0<x<\infty, \tag{4.1}$$

where τ and v satisfy

$$\frac{\partial\tau}{\partial t}-\frac{\partial v}{\partial x}=-\hat{\alpha}\tau,\qquad 0<x<\infty, \tag{4.2}$$

$$\frac{\partial v}{\partial t}-\frac{\partial\tau}{\partial x}=0,\qquad 0<x<\infty, \tag{4.3}$$

and

$$\hat{\alpha}=\begin{cases}1 & \text{if } \tau^2>\tau_y^2,\\ 1 & \text{if } \tau^2=\tau_y^2 \text{ and } \frac{\tau}{\tau_y^2}\frac{\partial v}{\partial x}>1,\\ \frac{\tau}{\tau_y^2}\frac{\partial v}{\partial x}, & \text{if } \tau^2=\tau_y^2 \text{ and } 0\leq\frac{\tau}{\tau_y^2}\frac{\partial v}{\partial x}\leq 1,\\ 0 & \text{if } \tau^2=\tau_y^2 \text{ and } \frac{\tau}{\tau_y^2}\frac{\partial v}{\partial x}<0,\\ 0 & \text{if } \tau^2<\tau_y^2,\end{cases} \tag{4.4}$$

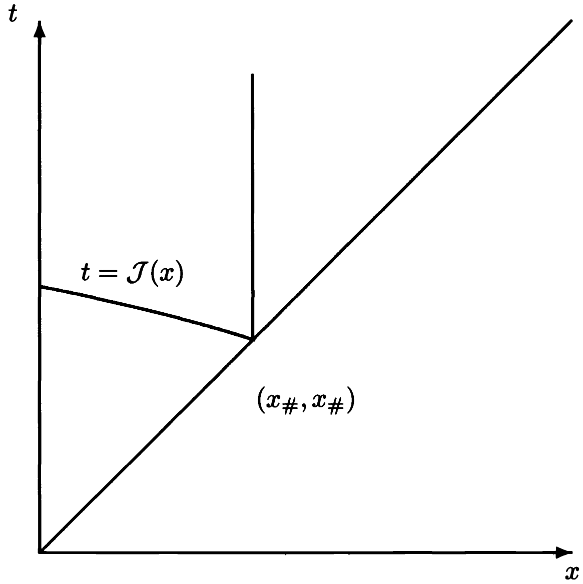


FIG. 1

and the initial and boundary conditions

$$(\tau, v)(x,0) = (0,0), \qquad 0 < x < \infty \tag{4.5}$$

and

$$v(0,t) = -\tau_0, \quad \text{where } \tau_0 > \tau_y. \tag{4.6}$$

We note that the results of the previous section guarantee there is at most one solution to the above problem.

In the region $0 \le t < x$, we have $(\tau, v) \equiv (0,0)$. Moreover, $\tau + v$ is continuous across the curve $t = x$ and thus satisfies $\tau_-(t,t) + v_-(t,t) \equiv 0$. The difficult part of the problem is to show there is a curve $t = \mathcal{J}(x)$, $0 < x < x_{\#}$, with $-1 < d\mathcal{J}/dx \le 0$, such that in the region $x < t < \mathcal{J}(x)$ with $0 < x < x_{\#}$, τ and v satisfy

$$\tau > \tau_y, \tag{4.7}$$

$$\frac{\partial \tau}{\partial t} - \frac{\partial v}{\partial x} = -\tau \quad \text{and} \quad \frac{\partial v}{\partial t} - \frac{\partial \tau}{\partial x} = 0, \tag{4.8}$$

the boundary condition (4.6) and $\tau_-(t,t) + v_-(t,t) = 0$. On the curve $t = \mathcal{J}(x)$ we have $\lim_{\epsilon \to 0^+} \tau(x, \mathcal{J}(x) - \epsilon) = \tau_y$ and $\hat{v}(x) \overset{\text{def}}{:=} \lim_{\epsilon \to 0^+} v(x, \mathcal{J}(x) - \epsilon)$ satisfies $0 \le d\hat{v}/dx \le \tau_y$. In the region $\mathcal{J}(x) < t$ and $0 < x < x_{\#}$ we have $\tau(x,t) = \tau_y$ and $v(x,t) = \hat{v}(x)$, whereas in $x_{\#} \le x < t$, $\tau \equiv \tau_y$ and $v(x,t) = \hat{v}(x_{\#}) = -\tau_y$ (see Fig. 1).

The existence of a curve $t = \mathcal{J}(x)$ with the desired properties may be established by converting the system (4.6), (4.8), and $\tau_-(t,t) + v_-(t,t) = 0$ to integral equations for τ and v in $x < t$, verifying that for $0 < t - x \ll 1$ the stress satisfies $\tau > \tau_y$, and finally by obtaining qualitative information on the level line $t = \mathcal{J}(x)$ defined by $\lim_{\epsilon \to 0^+} \tau(x, \mathcal{J}(x) - \epsilon) = \tau_y$.

Rather than perusing that approach we shall show how to obtain simple approximate solutions satisfying (4.6)–(4.8) and $\tau_-(t,t) + v_-(t,t) = 0$ as well as approximations to the level line $t = \mathcal{J}(x)$.

We note that for each integer $N \geq 1$ the system (4.8) has solutions

(4.9)
$$v_N = -\tau_0 + \sum_{k=1}^{N} \lambda_{(k)}(t) x^{2k-1}$$

and

(4.10)[6]
$$\tau_N = \lambda_0(t) + \sum_{k=1}^{N} \dot{\lambda}_{(k)} x^{2k}/2k$$

where the coefficients satisfy

(4.11)
$$\dot{\lambda}_0 + \lambda_0 = \lambda_1,$$

(4.12)
$$\ddot{\lambda}_k + \dot{\lambda}_k = 2k(2k+1)\lambda_{k+1}, \qquad 1 \leq k \leq N-1,$$

and

(4.13)
$$\ddot{\lambda}_N + \dot{\lambda}_N = 0.$$

These solutions satisfy the boundary condition $v_N(0^+, t) = -\tau_0$ and have $2N+1$ free parameters which are determined by insisting that the equation

(4.14)
$$\tau_N(t,t) + v_N(t,t) = 0$$

is satisfied to $O(t^{2N})$ as $t \to 0^+$. The approximate curve $t = \mathcal{J}_N(x)$ is subsequently determined by solving $\tau_N(x, \mathcal{J}_N(x)) = \tau_y$. An easy calculation shows that $\mathcal{J}_N(x) = O((\tau_0 - \tau_y)/\tau_0)$ and $d\mathcal{J}_N/dx < 0$ which guarantees that the number $x^N_{\#}$ defined by $\mathcal{J}_N(x^N_{\#}) = x^N_{\#}$ is $O((\tau_0 - \tau_y)/\tau_0)$, and thus on the boundary $x = t$, $\tau_N(t,t) + v_N(t,t)$ is at worst $O((\tau_0 - \tau_y)/\tau_0)^{2N+1}$ for $0 \leq t \leq x^N_{\#}$. We continue the approximate solutions to the rest of the region described by Fig. 1 via the extensions procedure used for the exact solution; that is, for $0 < t < x$,

(4.15)
$$(\tau_N, v_N) \equiv (0, 0),$$

for $\mathcal{J}_N(x) < t$ and $0 < x < x^N_{\#}$

(4.16)
$$v_N(x,t) = v_N(x, \mathcal{J}_N(x)) \quad \text{and} \quad \tau_N(x,t) = \tau_y,$$

and for $x_{\#} \leq x < t$,

(4.17)
$$v_N(x,t) = v_N(x_{\#}, \mathcal{J}_N(x_{\#})) \quad \text{and} \quad \tau_N(x,t) = \tau_y.$$

We are then guaranteed that the error made in failing to meet the boundary condition $\tau_N(t,t) + v_N(t,t) = 0$ is at worst $O((\tau_0 - \tau_y)/\tau_0)^{2N+1}$ for all $t > 0$. We shall present the details of this procedure for the case $N = 1$.

In this case,

(4.18)
$$v_1 = -\tau_0 + (\lambda_{1,0} + \lambda_{1,1} e^{-t}) x$$

[6] Here · denotes differentiation with respect to t.

and

$$\tau_1 = (\lambda_{1,0} + \lambda_{1,1} t e^{-t} + \lambda_{0,1} e^{-t}) - \frac{\lambda_{1,1} e^{-t} x^2}{2}, \tag{4.19}$$

and the insistence that $\tau_1(t,t) + v_1(t,t) = O(t^3)$ as $t \to 0^+$ implies that

$$\begin{aligned} \lambda_{1,0} + \lambda_{0,1} &= \tau_0 \\ \lambda_{1,0} - \lambda_{0,1} + 2\lambda_{1,1} &= 0 \\ \lambda_{0,1} - 5\lambda_{1,1} &= 0 \end{aligned} \tag{4.20}$$

and hence that

$$v_1 = \tau_0 \left(-1 + \frac{(3 + 5e^{-t})x}{8} \right) \tag{4.21}$$

and

$$\tau_1 = \frac{\tau_0}{8} \left(3 + t e^{-t} + 5e^{-t} - \frac{e^{-t} x^2}{2} \right). \tag{4.22}$$

The approximate curve $t = \mathcal{J}_1(x)$ is obtained by solving $\tau_1(x, \mathcal{J}_1(x)) = \tau_y$ or equivalently the equation

$$\left(3 + \mathcal{J}_1 e^{-\mathcal{J}_1} + 5e^{-\mathcal{J}_1} - \frac{x^2 e^{-\mathcal{J}_1}}{2} \right) = \frac{8\tau_y}{\tau_0}. \tag{4.23}$$

The fact that $0 < \tau_y/\tau_0 < 1$ guarantees the unique solvability of this equation for $0 \leq x \ll 1$ and that $\mathcal{J}_1(0) = O(2((\tau_0 - \tau_y)/\tau_0))$. A quick calculation also shows that

$$\frac{d\mathcal{J}_1}{dx} = \frac{-2x}{(8 + 2\mathcal{J}_1 - x^2)} < 0. \tag{4.24}$$

The number $x^1_{\#}$, where $\mathcal{J}_1(x^1_{\#}) = x^1_{\#}$ satisfies

$$\left(3 + x^1_{\#} e^{-x^1_{\#}} + 5e^{-x^1_{\#}} - \frac{(x^1_{\#})^2 e^{-x^1_{\#}}}{2} \right) = \frac{8\tau_y}{\tau_0}, \tag{4.25}$$

and for $0 < \tau_0 - \tau_y$ small enough we are guaranteed that $x^1_{\#} = O((\tau_0 - \tau_y)/\tau_0)$. This estimate, when combined with (4.24), implies that $-1 < d\mathcal{J}_1/dx$ for $0 \leq x \leq x^1_{\#}$.

Our final task is to show that the function

$$\hat{v}_1(x) \overset{\text{def}}{:=} \tau_0 \left(-1 + \frac{(3 + 5e^{-\mathcal{J}_1(x)})x}{8} \right) \tag{4.26}$$

satisfies

$$0 \leq \frac{d\hat{v}_1}{dx}(x) \leq \tau_y, \qquad 0 \leq x \leq x^1_{\#}. \tag{4.27}$$

The defining relation (4.26) implies that

$$\frac{d\hat{v}_1}{dx}(x) = \tau_0 \left(\frac{3 + 5e^{-\mathcal{J}_1(x)}}{8} \right) - \frac{5\tau_0 e^{-\mathcal{J}_1(x)}}{8} x \mathcal{J}_1'(x), \tag{4.28}$$

and this relationship, when combined with (4.23) and (4.24), implies that

$$(4.29)\qquad \frac{d\hat{v}_1}{dx}(x) = \tau_y + \frac{\tau_0 e^{-\mathcal{J}_1}((x^2-2\mathcal{J}_1)(8+2\mathcal{J}_1-x^2)+20x^2)}{16(8+2\mathcal{J}_1-x^2)}.$$

The fact that $\mathcal{J}_1(x) \geq x^1_{\#}$ for $0 \leq x \leq x^1_{\#} = O((\tau_0-\tau_y)/\tau_0)$ implies that the second term in (4.29) is negative and this provides the desired upper bound for $d\hat{v}_1/dx$. The desired lower bound is an immediate consequence of (4.28) and the bounds for $d\mathcal{J}_1/dx$.

We conclude this section by contrasting the above solution with what obtains if we replace our flow rule—α given by (4.4)—with the one generated by (2.31) and (2.32) and also by the flow rule associated with a rate independent elastic-perfectly plastic material. In the former case, (4.2) is replaced by

$$(4.30)\qquad \frac{\partial \tau}{\partial t} - \frac{\partial v}{\partial x} = -\alpha\tau_y,$$

and (4.4) is unchanged.

In the region $0 \leq t < x$ we have $(\tau, v) \equiv (0,0)$, and $\tau + v$ is continuous across $x = t$. For $0 \leq x \leq t \leq 2(\tau_0-\tau_y)/\tau_y$ we have

$$(4.31)\qquad v = -\tau_0 + \frac{\tau_y x}{2} \quad\text{and}\quad \tau = \tau_0 - \frac{\tau_y t}{2};$$

for $0 \leq x \leq 2(\tau_0-\tau_y)/\tau_y$ and $t \geq 2(\tau_0-\tau_y)/\tau_y$ we have

$$(4.32)\qquad v = -\tau_0 + \frac{\tau_y x}{2} \quad\text{and}\quad \tau = \tau_y,$$

and finally for $2(\tau_0-\tau_y)/\tau_y \leq x < t$ we have

$$(4.33)\qquad v = -\tau_y \quad\text{and}\quad \tau = \tau_y.$$

With this flow rule the curve $t = \mathcal{J}(\cdot)$ is the constant function $\mathcal{J}(x) = 2(\tau_0 - \tau_y)/\tau_y$, $0 \leq x \leq 2(\tau_0-\tau_y)/\tau_y$. Equations (2.52) and (2.53), the initial conditions $(F_{31}, p_{31})(x,0) = (0,0)$ for $x > 0$, and (4.31)–(4.33) allow us to determine (F_{31}, p_{31}). The result is

$$(4.34)\quad (F_{31}, p_{31}) = \begin{cases} (0,0), \qquad 0 \leq t < x, \\ (\tau_0 + \tau_y(\frac{t}{2} - x), \tau_y(t-x)), \qquad 0 \leq x < t < \frac{2(\tau_0-\tau_y)}{\tau_y}, \\ (\tau_0 + \tau_y(\frac{t}{2} - x), \tau_{01} - \tau_y + \tau_y(\frac{t}{2} - x)), \frac{2(\tau_0-\tau_y)}{\tau_y} \leq t \quad\text{and} \\ 0 \leq x < \frac{2(\tau_0-\tau_y)}{\tau_y}, \\ (\tau_y, 0), \frac{2(\tau_0-\tau_y)}{\tau_y} < x < t. \end{cases}$$

It is worth noting that the above solution is unique. This can be established using the arguments of §3 directly on the system (4.30) and (4.3)–(4.6).

We now examine the signaling problem for a rate independent elastic-perfectly plastic material. Equations (4.1)–(4.3) and (4.5) and (4.6) still hold, except now $\hat{\alpha}$ is given by

$$(4.35)\qquad \hat{\alpha} = \begin{cases} 0 & \text{if } \tau^2 < \tau_y^2 \\ 0 & \text{if } \tau^2 = \tau_y^2 \quad\text{and}\quad \frac{\tau}{\tau_y^2}\frac{\partial v}{\partial x} < 0, \\ \frac{\tau}{\tau_y^2}\frac{\partial v}{\partial x} & \text{if } \tau^2 = \tau_y^2 \quad\text{and}\quad 0 \leq \frac{\tau}{\tau_y^2}\frac{\partial v}{\partial x}. \end{cases}$$

We also have

$$\frac{\partial F_{31}}{\partial t} - \frac{\partial v}{\partial x} = 0, \quad \frac{\partial p_{31}}{\partial t} = \hat{\alpha}\tau, \quad \text{and} \quad F_{31} = \tau + p_{31}, \tag{4.36}$$

and these satisfy the initial conditions

$$(F_{31}, p_{31})(x, 0) = (0, 0), x > 0. \tag{4.37}$$

We seek solutions with structure similar to that obtained for the previous two models. Specifically, a shock curve $t = \hat{t}(x)$ such that in the region $0 < t < \hat{t}(x)$,

$$(F_{31}, p_{31}, \tau, v) = (0, 0, 0, 0), \tag{4.38}$$

and in the region $t > \hat{t}(x)$ the shear stress τ is at yield, i.e.,

$$\tau(x, t) = \tau_y, \qquad \hat{t}(x) < t. \tag{4.39}$$

We interpret (4.3) and $(4.36)_1$ as conservation laws, and this, together with (4.38) and (4.39), implies that on $t = \hat{t}(x)$,

$$v^-(x, \hat{t}(x)) + \tau_y \frac{d\hat{t}}{dx} = 0 \tag{4.40}$$

and

$$F_{31}^-(x, \hat{t}(x)) + v^-(x, \hat{t}(x)) \frac{d\hat{t}}{dx} = 0. \tag{4.41}$$

Here, $(v^-, F_{31}^-)(x, \hat{t}(x)) = \lim_{\epsilon \to 0^+}(v, F_{31})(x - \epsilon, \hat{t}(x))$. The identity (4.39) also implies that in $t > \hat{t}(x)$ the velocity v is a function of x only. Near $x = 0$ we choose

$$v(x, t) = -\tau_0 + \lambda x, \lambda > 0. \tag{4.42}$$

With this choice we obtain

$$p_{31} = \lambda(t - \hat{t}(x)) + p_-(x) \tag{4.43}$$

and

$$F_{31} = \tau_y + \lambda(t - \hat{t}(x)) + p_-(x). \tag{4.44}$$

Equation (4.40), together with $\hat{t}(0) = 0$, then yields

$$\hat{t}(x) = \frac{\tau_0^2 - (\tau_0 - \lambda x)^2}{2\lambda\tau_y}, \tag{4.45}$$

and (4.41), (4.44), and (4.45) imply that

$$p_-(x) = \frac{(\tau_0 - \lambda x)^2 - \tau_y^2}{\tau_y}. \tag{4.46}$$

We now let

$$x_\# = \frac{\tau_0 - \tau_y}{\lambda} \tag{4.47}$$

and note that

$$p_-(x) > 0, \qquad 0 \le x < x_\#, \tag{4.48}$$

$$p_-(x_\#) = 0, \tag{4.49}$$

and

$$\frac{d\hat{t}}{dx}(x_\#) = 1. \tag{4.50}$$

In the region $(\tau_0^2 - (\tau_0 - \lambda x)^2)/2\lambda\tau_y < t$ and $0 \le x < x_\# = (\tau_0 - \tau_y)/\lambda$ our solution is given by

$$F_{31} = \tau_y + \lambda\left(t + \frac{\tau_0^2 - (\tau_0 - \lambda x)^2}{2\lambda\tau_y}\right), \tag{4.51}$$

$$p_{31} = \lambda\left(t + \frac{\tau_0^2 - (\tau_0 - \lambda x)^2}{2\lambda\tau_y}\right), \tag{4.52}$$

$$v = -\tau_0 + \lambda x, \tag{4.53}$$

$$\tau = \tau_y. \tag{4.54}$$

The shock curve is continued to $x > x_\#$ by

$$\hat{t}(x) = \frac{\tau_0^2 - \tau_y^2}{2\lambda\tau_y} + \left(x - \frac{\tau_0 - \tau_y}{\lambda}\right) \tag{4.55}$$

and in the region $((\tau_0^2 - \tau_y^2)/2\lambda\tau_y) + (x - ((\tau_0 - \tau_y)/\lambda)) < t$ and $(\tau_0 - \tau_y)/\lambda = x_\# < x$,

$$F_{31} = \tau_y, \quad p_{31} = 0, \quad v = -\tau_y, \quad \text{and } \tau = \tau_y. \tag{4.56}$$

The line $x = x_\# = (\tau_0 - \tau_y)/\lambda$ is a stationary contact discontinuity and across it p_{31} jumps while the other fields are continuous. The interesting fact about the signaling problem for this model is the lack of uniqueness of solutions; we have a compatible solution for every $\lambda > 0$. This observation points out one of the weaknesses of the classical model.

5. Computational experiments. In this section we present the results of a computational experiment performed on the normalized system (2.48)–(2.52) when the pressure gradient is zero. The results reported deal with a two-dimensional generalization of the signalling problem of the previous section.

The experiment deals with the system (2.48)–(2.51) solved in the region $r > 0$ and $\pi/2 < \theta < 2\pi$, where $r = \sqrt{x^2 + y^2}$. At time $t = 0$ we assume that

$$(\tau_{31}, \tau_{32}, v) = (0, 0, 0) \tag{5.1}$$

for $r > 0$ and $\pi/2 < \theta < 2\pi$, and for $t > 0$ we assume that

$$v\left(r, \frac{\pi}{2}^+\right) = v(r, 2\pi^-) = \tau_0, \qquad r > 0, \tag{5.2}$$

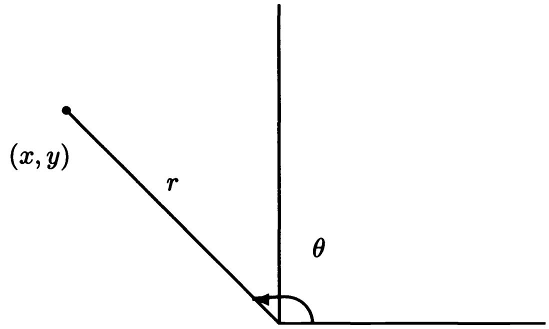


FIG. 2

where $\tau_0 > \tau_y$, and again $\tau_y > 0$ is the yield stress.

The elastic version of this problem, namely, the system

$$\frac{\partial \tau_{31}}{\partial t} - \frac{\partial v}{\partial x} = 0, \tag{5.3}$$

$$\frac{\partial \tau_{32}}{\partial t} - \frac{\partial v}{\partial y} = 0, \tag{5.4}$$

$$\frac{\partial v}{\partial t} - \frac{\partial \tau_{31}}{\partial x} - \frac{\partial \tau_{32}}{\partial y} = 0, \tag{5.5}$$

together with (5.1) and (5.2), was considered by Keller and Blank [13]. They obtained exact solutions to this and a number of other problems with self similar structure. Relevant to us here is the singular nature of $\tau_{31}^2 + \tau_{32}^2$ as $r \to 0^+$. Their results demonstrate that

$$\tau_{31}^2 + \tau_{32}^2 = O\left(\frac{t}{r}\right)^{2/3}, \qquad r \to 0^+. \tag{5.6}$$

This singular behavior also obtains for the plastic flow problem and forces us to treat the boundary conditions in our numerical simulation carefully. Our integration scheme for (2.48)–(2.51) is based on a symmetrized operator splitting algorithm for the governing differential equations. At time $t = nh$, $n = 0, 1, 2, \ldots$, our approximate solution consists of lattice data

$$(\tau_{31}, \tau_{32}, v)^n_{(k,m)} = (\tau_{31}, \tau_{32}, v)\left(\frac{(2k-1)}{2}h, \frac{(2m-1)}{2}h, nh\right). \tag{5.7}$$

For the problem under consideration the boundaries are not part of the computational lattice but are offset from it by a distance of $h/2$. The computational lattice is

$$\mathcal{S} = \{(k,m) \mid k \leq 0 \text{ and } m = 0, \pm 1, \pm 2, \ldots\} \cup \{(k,m) \mid k \geq 1 \text{ and } m \leq 0\}. \tag{5.8}$$

To update the data (5.7) we successively solve

$$\frac{\partial \tau_{31}}{\partial t} - \frac{\partial v}{\partial x} = 0, \quad \frac{\partial \tau_{32}}{\partial t} = 0, \quad \text{and} \quad \frac{\partial v}{\partial t} - \frac{\partial \tau_{31}}{\partial x} = 0, \quad 0 \leq t \leq h, \tag{5.9}$$

$$\frac{\partial\tau_{31}}{\partial t}=0,\quad \frac{\partial\tau_{32}}{\partial t}-\frac{\partial v}{\partial y}=0,\quad\text{and}\quad \frac{\partial v}{\partial t}-\frac{\partial\tau_{32}}{\partial y}=0,\quad 0\le t\le h, \tag{5.10}$$

and

$$\frac{\partial\tau_{31}}{\partial t}=-\hat\alpha\tau_{31},\quad \frac{\partial\tau_{32}}{\partial t}=-\hat\alpha\tau_{32},\quad\text{and}\quad \frac{\partial v}{\partial t}=0,\quad 0\le t\le h, \tag{5.11}$$

where of course $\hat\alpha$ is defined in (2.51). For (5.9) we use the approximate solution defined by (5.7) as initial data and let $(\tau_{31}^1,\tau_{32}^1,v^1)_{(k,m)}$ denote the value of this solution at $t=h$ on the lattice $\mathcal{S}$. We then solve (5.10) using the $(\tau_{31}^1,\tau_{32}^1,v^1)_{(k,m)}$ as initial data and let $(\tau_{31}^2,\tau_{32}^2,v^2)_{(k,m)}$ denote value of the solution at $t=h$ on $\mathcal{S}$. Finally, we solve (5.11) with $(\tau_{31}^2,\tau_{32}^2,v^2)_{(k,m)}$ as initial data and let $(\tau_{31}^3,\tau_{32}^3,v^3)_{(k,m)}$ denote the value of this solution at $t=h$ on $\mathcal{S}$.

We then repeat the process solving (5.10) first with the data (5.7), and, we let $(\tau_{31}^4,\tau_{32}^4,v^4)_{(k,m)}$ denote the lattice update at $t=h$. We then solve (5.9) using $(\tau_{31}^4,\tau_{32}^4,v^4)_{(k,m)}$ as initial data and let $(\tau_{31}^5,\tau_{32}^5,v^5)_{(k,m)}$ denote the lattice update. Finally we solve (5.11) with data $(\tau_{31}^5,\tau_{32}^5,v^5)_{(k,m)}$ and let $(\tau_{31}^6,\tau_{32}^6,v^6)_{(k,m)}$ denote the lattice update at $t=h$. The desired approximate solution $(\tau_{31},\tau_{32},v)_{(k,m)}^{(n+1)}$ is then obtained by averaging $(\tau_{31}^3,\tau_{32}^3,v^3)_{(k,m)}$ and $(\tau_{31}^6,\tau_{32}^6,v^6)_{(k,m)}$; that is,

$$(\tau_{31},\tau_{32},v)_{(k,m)}^{(n+1)}=\tfrac{1}{2}(\tau_{31}^3+\tau_{31}^6,\tau_{32}^3+\tau_{32}^6,v^3+v^6)_{(k,m)}. \tag{5.12}$$

Of course, all of the intermediate updates are solved subject to the boundary conditions of the original problem. Here these boundary conditions manifest themselves as reflection conditions at those lattice points that are a distance $h/2$ away form the actual boundary. Formal accuracy could be maintained if we used either $(\tau_{31}^3,\tau_{32}^3,v^3)_{(k,m)}$ or $(\tau_{31}^6,\tau_{32}^6,v^6)_{(k,m)}$ for the updated approximate solution but either of these updates alone would, over time, tend to introduce asymmetries into the approximates not present in the actual solution. These asymmetries are removed with the algorithm employed.

The results of our experiment are shown in Figs. 3–7. Each snapshot shows two different representations of the velocity field and the total shear stress, namely the quantity $\sqrt{\tau_{31}^2+\tau_{32}^2}$. This simulation was run with $h=1/50$, $\tau_y=1$, and $\tau_0=1.3$. The contours on the velocity plots are spaced 0.1 apart and run from $v=0$ to $v=1.3$. The stress contours run from 1 to 3.2 in increments of 0.2. In these snapshots one sees not only the plane wave solutions of the previous section but also the effect of the corner singularity which are confined to the region $0\le r\le t$ and $\pi/2<\theta<2\pi$.

For comparison we have run the elastic version of this problem with the same boundary conditions and same values of h, τ_y, and τ_0. These results are shown in Figs. 8–12.

It should be noted that for both problems the velocity fields satisfy the additional condition

$$\lim_{r\to t^+} v(r,\theta,t)=0,\qquad \frac{\pi}{2}<\theta<2\pi \tag{5.13}$$

and that our numerical solutions meet this consistency condition automatically.

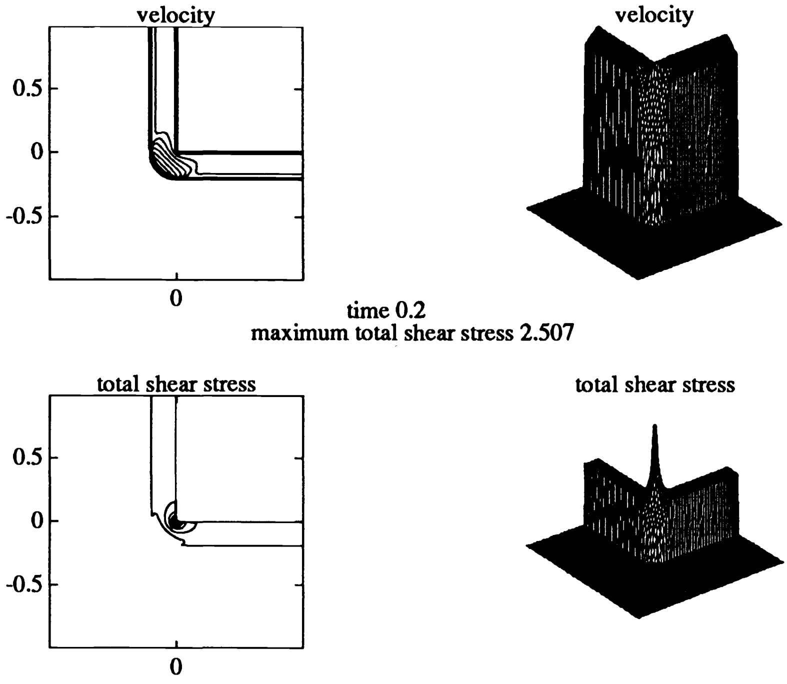

velocity
0.5
0
-0.5
0
velocity
time 0.2
maximum total shear stress 2.507
total shear stress
0.5
0
-0.5
0
total shear stress

FIG. 3

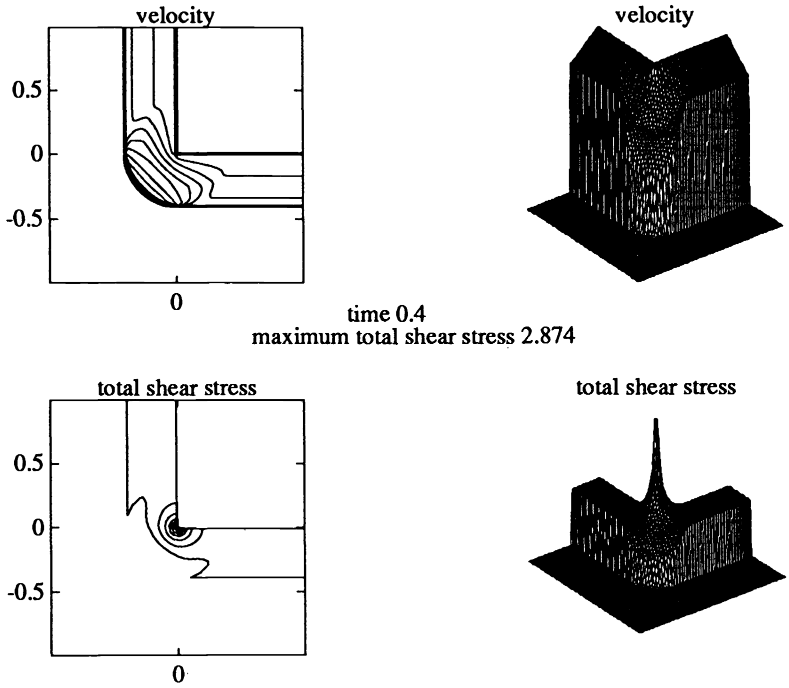
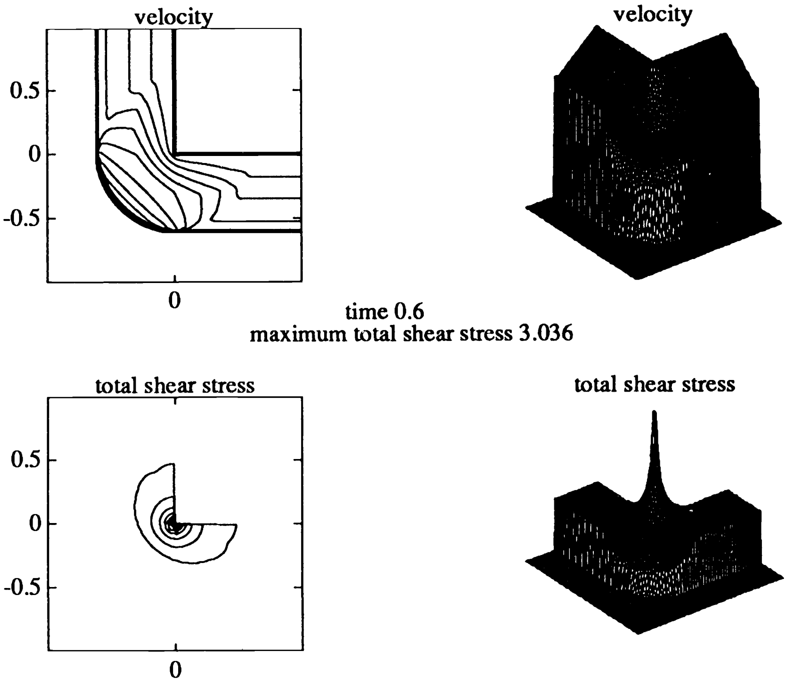
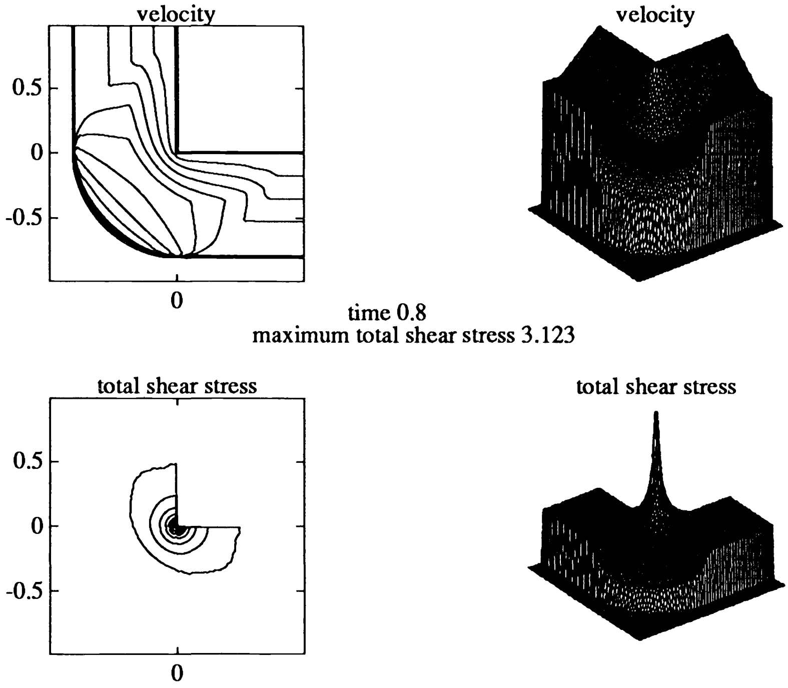

velocity
0.5
0
-0.5
0
velocity
time 0.4
maximum total shear stress 2.874
total shear stress
0.5
0
-0.5
0
total shear stress

FIG. 4



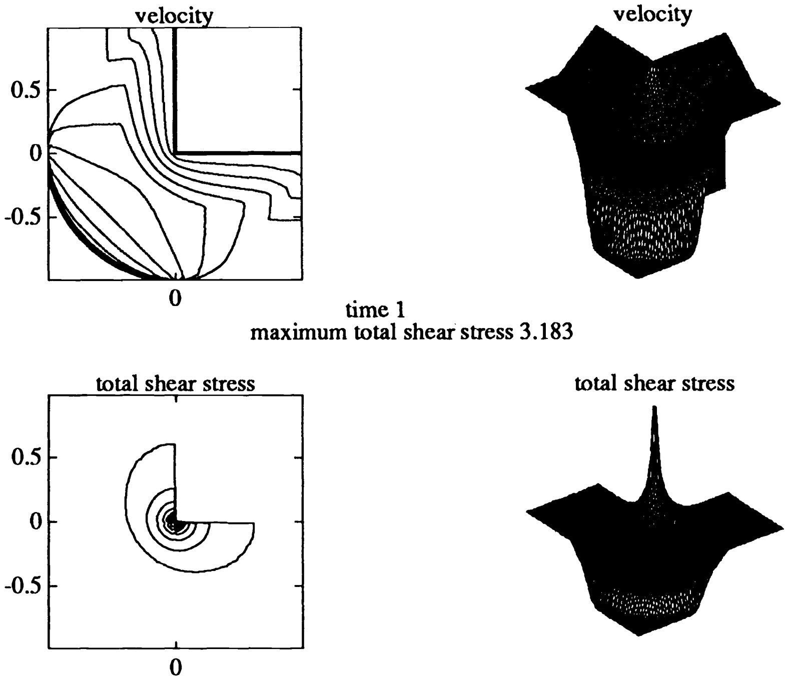
time 0.6
maximum total shear stress 3.036

FIG. 5



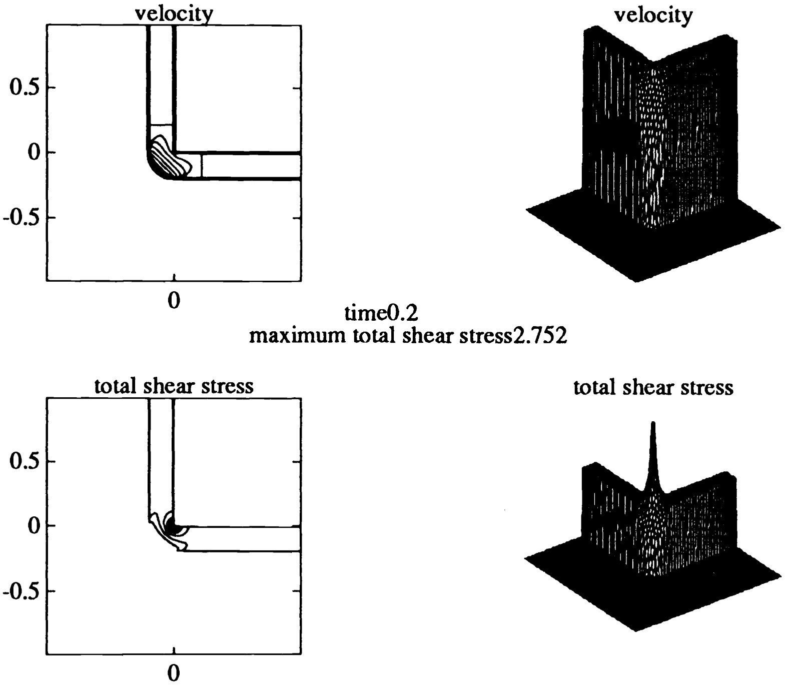
time 0.8
maximum total shear stress 3.123

FIG. 6



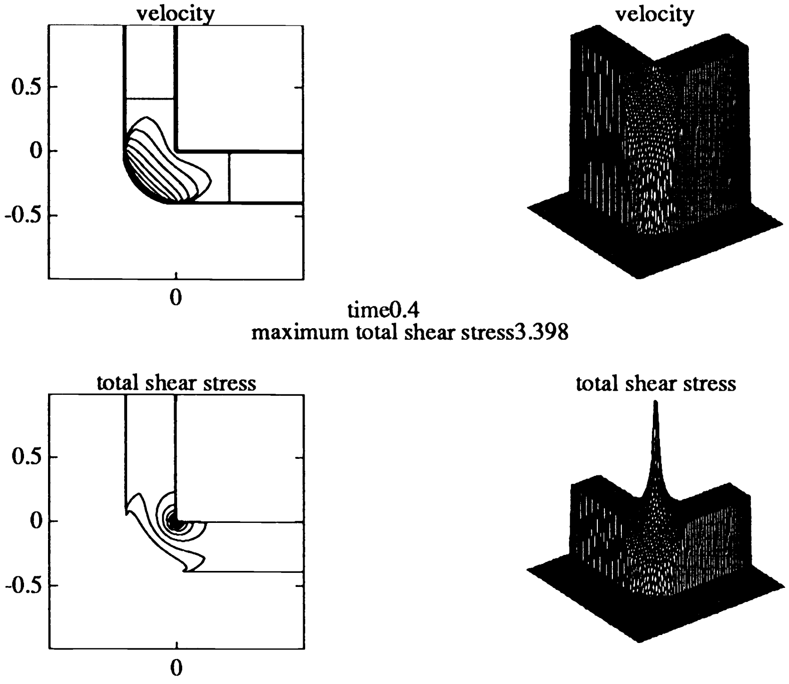
time 1
maximum total shear stress 3.183

FIG. 7



time0.2
maximum total shear stress2.752

FIG. 8



velocity

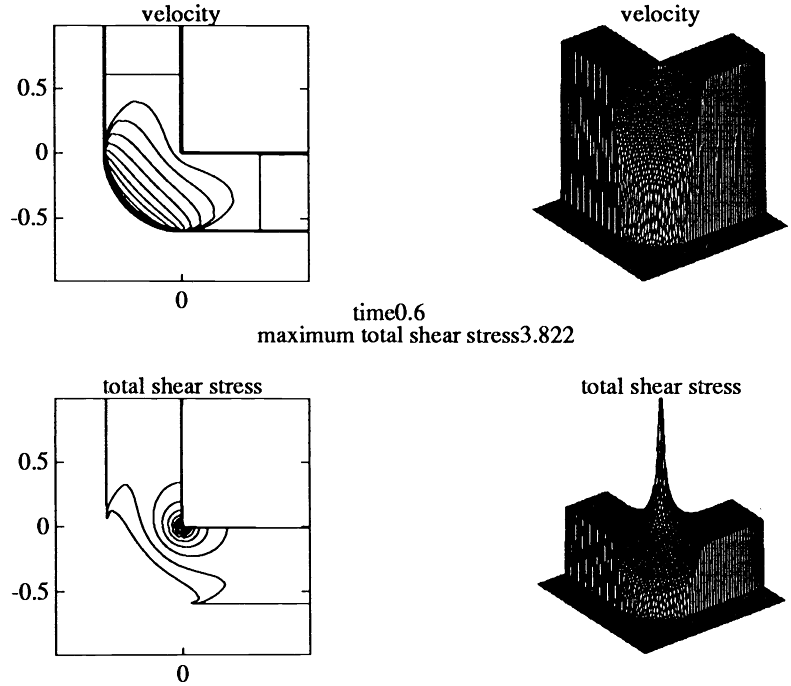
velocity

time0.4
maximum total shear stress3.398

total shear stress

total shear stress

FIG. 9



velocity

velocity

time0.6
maximum total shear stress3.822

total shear stress

total shear stress

FIG. 10

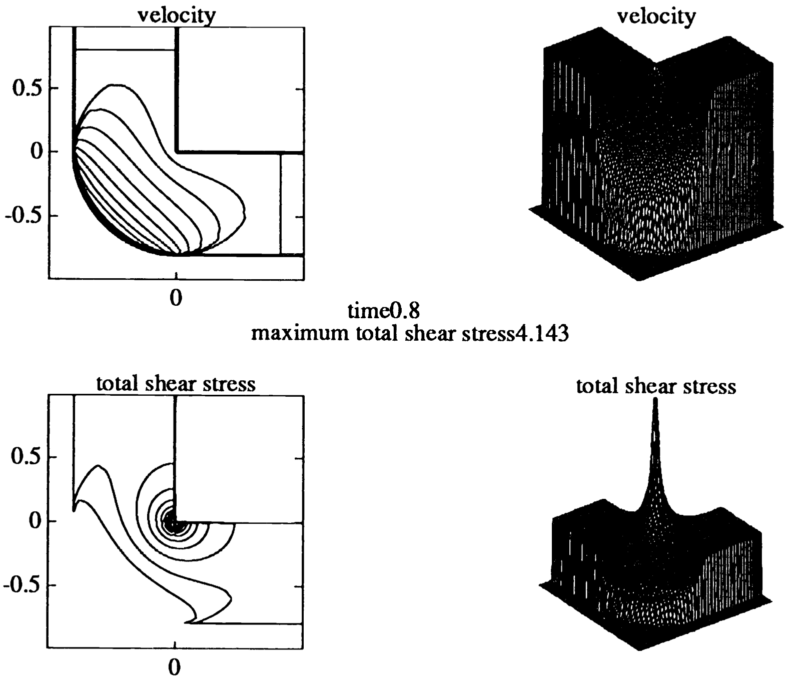

velocity
velocity
0.5
0
-0.5
0
time0.8
maximum total shear stress4.143
total shear stress
total shear stress
0.5
0
-0.5
0

FIG. 11

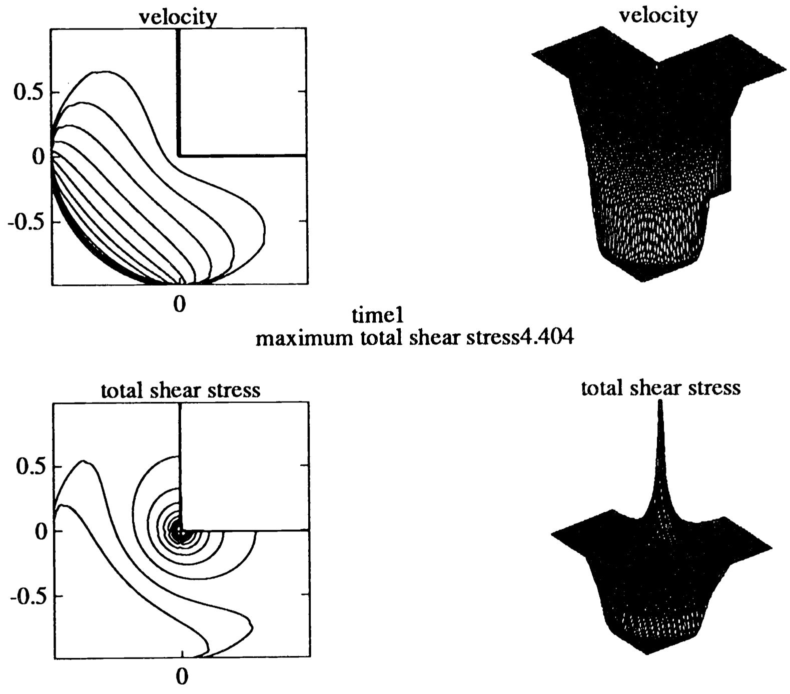

velocity
velocity
0.5
0
-0.5
0
time1
maximum total shear stress4.404
total shear stress
total shear stress
0.5
0
-0.5
0

FIG. 12

REFERENCES

[1] J. M. Greenberg, *Models of elastic-perfect plastic materials*, Euro. J. Appl. Math., 1 (1990), pp. 131–150.

[2] ———, *The longtime behavior of elastic-perfectly plastic materials*, in Free Boundary Problems Involving Solids—Proceedings of the International Colloquium Free Boundary Problems: Theory and Applications, J. M. Chadam and H. K. Rasmusseu, eds., Longman Scientific and Technical, Harlow, UK, 1993, pp. 28–34.

[3] T. I. Seidman, *Event-driven discontinuities for continuous-time systems*, in Proc. 28th IEEE Conf. Decision and Control, 1989, pp. 795–796.

[4] V. I. Utkin, *Sliding models and their application in variable structure systems*, Mir, Moscow, 1978.

[5] A. F. Filippov, *Differential equations with discontinuous righthand sides*, Kluwer, Dordrecht, 1978.

[6] S. S. Antman and W. G. Szymczak, *Nonlinear elastoplastic wave*, Contemp. Math., 100 (1989), pp. 27–54.

[7] ———, *Large antiplane shearing motion of nonlinear viscoplastic materials*, preprint, 1990.

[8] B. D. Coleman and D. R. Owen, *On Thermodynamics and Elastic–Plastic Materials*, Arch. Rational Mech. Anal., 59 (1975), pp. 25–51.

[9] J. L. Buhite and D. R. Owen, *An Ordinary Differential Equation From the Theory of Plasticity*, Arch. Rat Mech. Anal., 71 (1979), pp. 357–383.

[10] B. D. Coleman and M. L. Hodgdon, *On Shear Bands in Ductile Materials*, Arch. Rational Mech. Anal., 90 (1985), pp. 219–247.

[11] D. R. Owen, *Weakly decaying energy separation and uniqueness of motions of an elastic–plastic oscillator with work-hardening*, Arch. Rational Mech. Anal., 98 (1987), pp. 95–114.

[12] M. E. Gurtin, *Topics in Finite Elasticity*, CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1981.

[13] J. B. Keller and A. Blank, *Diffraction and reflection of pulses by wedges and corners*, Comm. Pure and Appl. Math., 4 (1951), pp. 75–91.

SIAM J. MATH. ANAL.
Vol. 24, No. 4, pp. 968–986, July 1993

A GEOMETRIC SINGULAR PERTURBATION ANALYSIS OF DETONATION AND DEFLAGRATION WAVES*

I. GASSER† AND P. SZMOLYAN†

Abstract. The existence of steady plane wave solutions of the Navier–Stokes equations for a reacting gas is analyzed. Under the assumption of an ignition temperature the existence of detonation and deflagration waves close to the corresponding waves of the ZND-model is proved in the limit of small viscosity, heat conductivity, and diffusion. The method is constructive, since the classical solutions of the ZND-model serve as singular solutions in the context of geometric singular perturbation theory. The singular solutions consist of orbits on which the dynamics are slow-driven by chemical reaction and of orbits on which the dynamics are fast-driven by gasdynamic shocks. The approach is geometric and leads to a clear, complete picture of the existence, structure, and asymptotic behavior of detonation and deflagration waves.

Key words. detonations, deflagrations, shock waves, traveling waves, singular perturbations

AMS subject classifications. 34C37, 34E15, 35L67, 76L05

1. Introduction.

We study plane wave solutions of the Navier–Stokes equations for a reacting gas in one space dimension. We consider the simplest possible chemical reaction of one reactant that is converted to the product by a one-step exothermic reaction. The equations governing the reacting flow are given by

$$\begin{aligned}(\rho)_t + (\rho u)_x &= 0,\\ (\rho u)_t + (\rho u^2 + p)_x &= (\mu u_x)_x,\\ \left[\rho\left(\frac{u^2}{2}+e\right)\right]_t + \left[\rho u\left(\frac{u^2}{2}+e\right)+pu\right]_x &= (\lambda T_x)_x + (q\rho D Y_x)_x + (\mu u u_x)_x,\\ (\rho Y)_t + (\rho u Y)_x &= (\rho D Y_x)_x - k\rho Y\phi(T).\end{aligned} \tag{1.1}$$

In these equations the variables ρ, u, p, e, T, and Y are the density, velocity, pressure, specific energy, temperature, and reactant mass fraction of the gas. The unburned gas corresponds to $Y = 1$; a totally burned gas corresponds to $Y = 0$. All variables will be made dimensionless later on. The constants μ, λ, and D are, respectively, viscosity, heat conductivity, and diffusion coefficients. We assume an ideal gas, i.e., the pressure is given by $p = R\rho T$, where R is the gas constant. The specific energy is given by $e = c_v T + qY$, where c_v is the specific heat at constant volume and q is the heat release parameter. The assumption of an exothermic reaction implies that q is positive. For the usual Arrhenius kinetics $\phi(T) = e^{-E/RT}$ holds, where E is an activation energy. We will have to modify this later because of the cold boundary difficulty. The above equations are standard; see, e.g., Fickett and Davis [3] or Williams [16].

Detonation and deflagration waves are traveling wave solutions, i.e., solutions depending only on $\xi = x - ct$, of (1.1) connecting an unburned state at $\xi = -\infty$ to a burned state at $\xi = \infty$. Due to Galilei invariance it suffices to consider the case $c = 0$, that is, the stationary problem corresponding to (1.1).

* Received by the editors January 29, 1992; accepted for publication (in revised form) September 3, 1992.

† Institut für Angewandte und Numerische Mathematik, Technische Universität Wien, Wiedner Hauptstraße 8-10, Austria.

$$
\begin{aligned}
(\rho u)_x &= 0, \\
(\rho u^2 + p)_x &= (\mu u_x)_x, \\
\left[\rho u\left(\frac{u^2}{2} + e\right) + pu\right]_x &= (\lambda T_x)_x + (q\rho D Y_x)_x + (\mu u u_x)_x, \\
(\rho u Y)_x - (\rho D Y_x)_x &= -k\rho Y\phi(T).
\end{aligned}
\tag{1.2}
$$

Throughout this paper we denote by f_-, respectively, f_+, the value of any function f at $x = -\infty$, respectively, $x = \infty$. Neglecting all dissipative effects, i.e., setting $\lambda = \mu = D = 0$, in (1.2) gives the ZND-model (named after Zeldovich, Neumann, and Döring):

$$
\begin{aligned}
(\rho u)_x &= 0, \\
(\rho u^2 + p)_x &= 0, \\
\left[\rho u\left(\frac{u^2}{2} + e\right) + pu\right]_x &= 0, \\
(\rho u Y)_x &= -k\rho Y\phi(T).
\end{aligned}
\tag{1.3}
$$

We briefly review the classical analysis of detonation and deflagration waves for the ZND-model; see, e.g., Courant and Friedrichs [1] and Williams [16], since it will be basic in our analysis of the corresponding waves of (1.2). The first equation in (1.3) implies that the mass flux ρu has a constant value denoted by $m = \rho_- u_-$. The fluxes of momentum and energy are also constant; this gives the Rankine–Hugoniot conditions for a shock wave, which describe possible burned states ρ and p for given ρ_-, p_-, and Y_-. By integrating the second and the third equation in (1.3) we obtain the equation of the Raleigh line

$$
p - p_- = -m^2\left(\frac{1}{\rho} - \frac{1}{\rho_-}\right)
\tag{1.4}
$$

and the equation of the Hugoniot curves

$$
e_- - e = \frac{1}{2}(p + p_-)\left(\frac{1}{\rho} - \frac{1}{\rho_-}\right).
\tag{1.5}
$$

Due to the dependence of the internal energy on the mass fraction Y of the reactant, the usual Hugoniot curve of gasdynamics is shifted for $Y \neq Y_-$. For a given state on the left there exist—depending on the value of m—two, one, or no completely burned right state. The burned state for the critical value of m, for which there exists just one burned state, is the Chapman–Jouget point. There are two fundamentally different processes possible; those that are compressive are called detonations, the ones that are expansive are called deflagrations. The burned state on the detonation branch closer to the unburned state is called weak detonation point; the corresponding process is called a weak detonation. The burned state on the detonation branch farther away from the unburned state is called strong detonation point; the corresponding process is called a strong detonation. A similar classification holds for deflagrations; see Fig. 1. In a weak deflagration wave the variable Y is determined by a scalar differential equation; the gasdynamic variables are in equilibrium following a curve parametrized by Y. Strong deflagrations are ruled out because of entropy considerations. On the other hand, the ZND-structure for a strong detonation is that of an inert gasdynamic

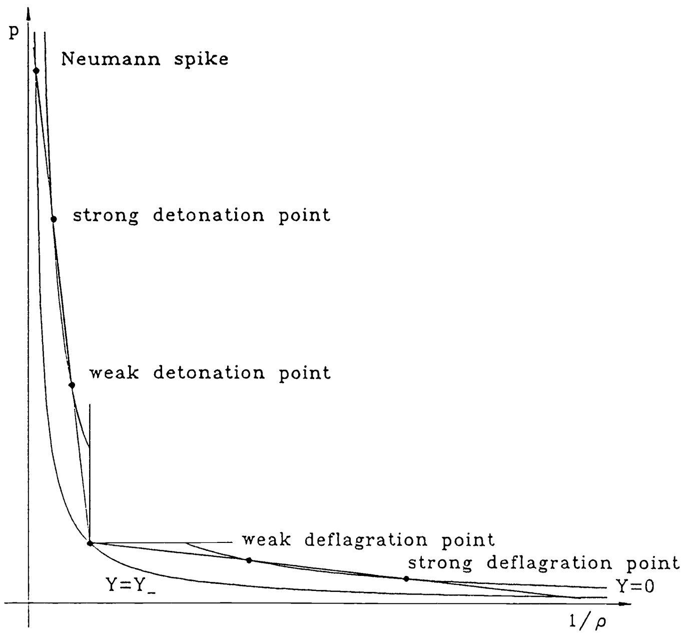


FIG. 1. *Chapman–Jouget diagram.*

shock wave that compresses and heats the gas, followed by a weak deflagration in which the reaction takes place. In Fig. 1 the ZND-structure of a strong detonation corresponds to an instantaneous jump from the unburned state to the intersection of the Raleigh line and the Hugoniot curve of the unburned state, followed by a deflagration along the Raleigh line to the burned state.

Clearly, there is some arbitrariness in this construction and questions of uniqueness arise; e.g., another possible structure of a strong detonation, starting at the unburned state as a gasdynamic shock followed by a deflagration to an intermediate partially burned state followed by another shock and another deflagration wave to the burned state, cannot be ruled out by considering (1.3) alone.

A standard criterion to distinguish unphysical solutions of hyperbolic conservation laws—obtained by neglecting all dissipative effects—from physical solutions is to accept only those solutions which are limits of solutions of the dissipative equations as the dissipation coefficients go to zero. For initial value problems corresponding to hyperbolic conservation laws this is one of the main open problems of the subject. On the level of shock waves, however, the existence of viscous profiles, which converge to the shock wave as the dissipation coefficients go to zero, is a well-established admissibility criterion (see, e.g., Smoller [12]).

Thus, the question of admissibility of ZND-waves leads to the question of the existence of solutions of (1.2) connecting an unburned state to a burned state. Additionally, the convergence of solutions of (1.2) to solutions of (1.3) for $(\lambda, \mu, D) \to 0$ has to be analysed. The first difficulty encountered is the well-known cold boundary difficulty, i.e., the unburned state is no stationary point of (1.2) for the usual Arrhenius kinetics $\phi(T) = \exp(-E/RT)$, unless $T_- = 0$, holds. The usual remedy is to modify the function $\phi(T)$, such that $\phi(T)$ vanishes identically for T below a certain ignition temperature T_i. More precise assumptions will be made later. Under this assumption detonation and deflagration waves are heteroclinic orbits connecting fixed points of (1.2).

The existence of detonation and deflagration waves of system (1.1) has been analysed by many authors, usually under the simplifying assumptions: Prandtl number $Pr = \frac{3}{4}$ and (or) Lewis number $Le = 1$. We refer to Hirschfelder and Curtiss [9] and Wood [17]. The activation energy E is typically very large, thus another frequently used approach is to consider (1.2) in the limit $E \to \infty$; see Lu and Ludford [11] and Holmes and Stewart [10]. Gardner [4] used the Conley index to prove the existence of detonation waves for the reactive Navier–Stokes equations in Lagrangian coordinates. He made no assumptions on Lewis and Prandtl numbers, but omitted the species diffusion term from the energy balance equation. The convergence of these solutions of (1.2) to solutions of the ZND-model is not discussed by these authors. Recently, Wagner [14], [15] obtained results on the ZND-limit for detonation and deflagration waves. Under the assumption of Prandtl number $\frac{3}{4}$ he proves existence of viscous profiles by topological arguments. A priori estimates and a compactness argument allow him to conclude that ZND-waves are (almost everywhere) limits of (certain subsequences of) solutions of (1.2) for $\lambda \to 0$ and constant ratios μ/λ and D/λ.

In our approach we only assume the existence of an ignition temperature and that the values of μ, λ, and D in a scaled version of (1.2) are small. No assumptions on the Lewis number and the Prandtl number are made. We consider system (1.2) as a singularly perturbed system of ordinary differential equations, which can be written in the form

$$\begin{aligned}
\mu u_x &= m(u-u_\pm) + mR\left(\frac{T}{u} - \frac{T_\pm}{u_\pm}\right),\\
\lambda T_x &= mc_p(T-T_\pm) - mRu\left(\frac{T}{u} - \frac{T_\pm}{u_\pm}\right) + mq(Z-Z_\pm) - \frac{m}{2}(u-u_\pm)^2,\\
DY_x &= u(Y-Z),\\
Z_x &= -k\frac{Y}{u}\phi(T),
\end{aligned} \tag{1.6}$$

where the variable Z is defined by the third equation (see §2). We prove the existence of detonation and deflagration waves close to the corresponding waves of the ZND-model. Our method is constructive since the classical solutions of the ZND-model serve as singular solutions in the context of geometric singular perturbation theory. The singular solutions consist of orbits on which the dynamics are slow-driven by the chemical reaction and of orbits on which the dynamics are fast-driven by gasdynamic shocks. Methods from dynamical systems theory allow us to conclude the existence of solutions of system (1.6) close to these singular solutions. Our approach is geometric and leads to a clear, complete picture of the existence, structure, and asymptotic behavior of detonation and deflagration waves. We obtain a complete characterization of the global flow in a two-dimensional invariant manifold, which contains all fixed points and all heteroclinic orbits of (1.6); from this point of view, the situation becomes particularly transparent. In the usual situation, that for a given unburned state (ρ_-, u_-, T_-, Y_-) the temperature at the two corresponding burned states is above ignition, our results can be summarized in the following theorem.

MAIN THEOREM. *Assume that the function* $\phi(T)$ *is smooth, with ignition temperature* T_i. *For* $\mu = \varepsilon\hat{\mu}$, $\lambda = \varepsilon\hat{\lambda}$, $D = \varepsilon\hat{D}$ *with* $\hat{\mu}, \hat{\lambda}, \hat{D} > 0$ *there exists* $\varepsilon_0 = \varepsilon_0(\hat{\mu}, \hat{\lambda}, \hat{D}) > 0$ *such that the following assertions hold for* $0 < \varepsilon < \varepsilon_0$.

(i) *Let* (ρ_-, u_-, T_-, Y_-), *with* $T_- < T_i$, *be an unburned state corresponding to a detonation. Then there exists a unique strong detonation wave solution of* (1.2).

(ii) *Let* (ρ_-, u_-, T_-, Y_-), *with* $T_- = T_i$, *be an unburned state corresponding to a detonation. Then there exists a two-dimensional manifold (with boundary) formed by all strong detonation wave solution of* (1.2). *The boundary of the manifold of strong detonation waves is formed by a strong detonation wave that has the appearance of a gasdynamic shock followed by a weak deflagration and by a unique weak detonation wave followed by a gasdynamic shock.*

(iii) *Let* (ρ_-, u_-, T_-, Y_-) *be an unburned state corresponding to a deflagration. Then there exists a unique weak deflagration wave solution of* (1.2) *if and only if* $T_- = T_i$ *holds. Strong deflagration wave solutions do not exist.*

(iv) *In suitable parametrizations all the detonation and deflagration waves converge to the corresponding waves of the* ZND*-model, uniformly in* C^1*-norm away from the location of gasdynamic shocks as* $\varepsilon \to 0$.

We remark that our method is not directly applicable in the case of Chapman–Jouget processes because the Chapman–Jouget point is a nonhyperbolic fixed point of (1.6) corresponding to a turning point in the context of singular perturbation theory.

We briefly outline the organization of the paper. In §2 we derive (1.6) and perform the scaling to obtain the final form of the governing equations. The necessary results from geometric singular perturbation theory are briefly explained in §3. In §4 we apply our method to the simpler problem of viscous profiles for gasdynamic shocks. This problem has been solved by Gilbarg [7]; however, we have included this example to illustrate the method and because we need the results on the structure of gasdynamic shocks in our analysis of detonation waves. The main result is proved in §5. Technical difficulties due to the ignition temperature assumption require a slight extension of the methods of §3. In §6 we discuss qualitative properties of detonation and deflagration waves, and prove the existence of incompletely burned detonation and deflagration waves.

2. Singularly perturbed scaled equations. By using $\rho u = m$ and integrating the second and third equation in (1.2) we obtain

$$\mu u_x = m(u - u_\pm) + mR\left(\frac{T}{u} - \frac{T_\pm}{u_\pm}\right), \tag{2.1}$$

$$\lambda T_x + \mu u u_x + q\rho D Y_x = m[c_v(T - T_\pm) + q(Y - Y_\pm) + (u^2 - u_\pm^2)/2 + R(T - T_\pm)]. \tag{2.2}$$

We define a new variable

$$Z = Y - \rho D\frac{Y_x}{m} = Y - D\frac{Y_x}{u} \tag{2.3}$$

to obtain a first-order system. Since Y_x must vanish at $x = \pm\infty$,

$$Z_- = Y_-, \qquad Z_+ = Y_+ \tag{2.4}$$

must hold. In the new variable the equation for Y in (1.2) has the form

$$Z_x = -k\frac{Y}{u}\phi(T). \tag{2.5}$$

By substituting for $\mu u u_x$ and $q\rho D Y_x$ from (2.1) and (2.3), and by using the relation

$$c_p = c_v + R \tag{2.6}$$

for the specific heat at constant pressure c_p and the specific heat at constant volume c_v, we obtain the final form of (2.2):

$$\text{(2.7)}\quad \lambda T_x = mc_p(T - T_\pm) - mRu\left(\frac{T}{u} - \frac{T_\pm}{u_\pm}\right) + mq(Y - Y_\pm) - \frac{m}{2}(u - u_\pm)^2.$$

Equations (2.1), (2.7), (2.3), and (2.5) are system (1.6). We denote by $M^2 = u^2/\gamma RT$ the square of the Mach number, where $\gamma = c_p/c_v$ is the ratio of the specific heats. For given $P_- = (u_-, T_-, Y_-, Z_-)$ we are interested in completely burned fixed points of (1.6). A straightforward calculation yields the following lemma.

LEMMA 2.1. *For a given unburned state P_-, there exist fixed points corresponding to completely burned states if and only if*

$$\text{(2.8)}\qquad \frac{(M_-^2 - 1)^2}{2(\gamma + 1)M_-^2} \geq qY_-$$

holds; $M_-^2 > 1$ corresponds to a detonation, and $M_-^2 < 1$ corresponds to a deflagration. Strict inequality in (2.8) *implies the existence of a completely burned state $P^\star = (u^\star, T^\star, 0, 0)$, with $M^{\star 2} < 1$, corresponding to a strong detonation or a weak deflagration, and of a state $P_\star = (u_\star, T_\star, 0, 0)$, with $M_\star^2 > 1$, corresponding to a weak detonation or a strong deflagration. Equality in* (2.8) *implies the existence of a unique completely burned state—with Mach number one—corresponding to the Chapman–Jouget detonation or deflagration point.*

In the following we set $(u_+, T_+, Y_+, Z_+) = (u^\star, T^\star, 0, 0)$ in (1.6) and use the scaling in Table 1 to make the equations dimensionless.

TABLE 1
Scaling.

Quantity	Unit	Scaling factor
x	m	$u^\star/k$
u	m s^{-1}	$u^\star$
T	K	$T^\star$
Z	1	
Y	1	
μ	kg m s^{-1}	$mu^\star/k$
λ	kg m^3 s^{-3} K^{-1}	$c_p mu^\star/k$
D	m^2 s^{-1}	$u^{\star 2}/k$
m	kg s^{-1}	
R, c_v, c_p	m^2 s^{-2} K^{-1}	
q	m^2 s^{-2}	$c_pT^\star$
k	s^{-1}	

The scaled quantity is obtained by dividing the unscaled quantity by its reference value, and is denoted by superscript $\sim$. If we define $\tilde\phi(\tilde T) = \phi(\tilde T T^\star)$, we obtain the scaled equations

$$\text{(2.9)}\qquad \tilde\mu \tilde u_{\tilde x} = \tilde u - 1 + \frac{1}{\gamma M^{\star 2}}\left(\frac{\tilde T}{\tilde u} - 1\right),$$

$$\text{(2.10)}\qquad \tilde\lambda \tilde T_{\tilde x} = \tilde T - 1 - \frac{\gamma - 1}{\gamma}(\tilde T - \tilde u) + \tilde q Z - \frac{(\gamma - 1)M^{\star 2}}{2}(\tilde u - 1)^2,$$

$$\text{(2.11)}\qquad \tilde D Y_{\tilde x} = \tilde u(Y - Z),$$

$$\text{(2.12)}\qquad Z_{\tilde x} = -\frac{Y}{\tilde u}\tilde\phi(\tilde T).$$

System (2.9)–(2.12) is singularly perturbed, because the parameters $\tilde{\mu}$, $\tilde{\lambda}$, and $\tilde{D}$ are typically very small. The actual size of these parameters in a given problem is given by the scaling in Table 1. In this paper we consider the ZND-limit, i.e., the limit of vanishing dissipation (see §1). In the limit process $(\tilde{\mu}, \tilde{\lambda}, \tilde{D}) \rightarrow 0$ various singular limits are possible depending on the ratios of the parameters. Typically, the Lewis number $Le \sim \tilde{\lambda}/\tilde{D}$ and the Prandtl number $Pr = \tilde{\mu}/\tilde{\lambda}$ satisfy $Le \sim 1$ and $Pr \sim \frac{3}{4}$ (see [16]). Therefore, it is reasonable to consider the limit $(\tilde{\mu}, \tilde{\lambda}, \tilde{D}) \rightarrow 0$ with fixed ratios Pr and Le. However, we do not make the common simplifying assumptions $Le = 1$ and $Pr = \frac{3}{4}$. The analysis of the ZND-limit for the case of constant ratios Pr and Le is carried out in §5, certain extensions to nonconstant ratios and other singular limits are briefly discussed in §6.

From here on we use the scaled quantities, therefore, we drop the superscript $\tilde{}$. The scaled fixed points are still denoted by $P^\star$, $P_\star$, and P_-; obviously, $P^\star = (1, 1, 0, 0)$ holds.

3. Geometric singular perturbation theory. The dynamical systems approach to singular perturbation problems origins—in its modern form—in the work of Fenichel [2], but has only recently become more popular. In Szmolyan [13] a method—based on this invariant manifold approach—is developed to prove the existence of transversal heteroclinic orbits of singularly perturbed differential equations. In this section we briefly summarize the necessary results from [2], [13] to provide the framework for our analysis of (2.9)–(2.12) in §§4 and 5. We consider singularly perturbed systems of differential equations in the standard form

$$\begin{aligned} \dot{x} &= f(x, y), \\ \varepsilon \dot{y} &= g(x, y), \end{aligned} \tag{3.1}$$

with $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$, $\varepsilon_0 > 0$ small, and $(x, y) \in \mathbf{R}^{m+k}$. We assume that $f \in \mathbf{R}^m$ and $g \in \mathbf{R}^k$ are C^r functions of (x, y), with $r \geq 2$. The independent variable is t. We call (3.1) the slow problem. By transforming to the variable $\tau = t/\varepsilon$ we obtain the equivalent fast problem

$$\begin{aligned} x' &= \varepsilon f(x, y), \\ y' &= g(x, y). \end{aligned} \tag{3.2}$$

By setting $\varepsilon = 0$ in (3.1) and (3.2) we obtain the reduced problem

$$\begin{aligned} \dot{x} &= f(x, y), \\ 0 &= g(x, y), \end{aligned} \tag{3.3}$$

and the layer problem

$$\begin{aligned} x' &= 0, \\ y' &= g(x, y). \end{aligned} \tag{3.4}$$

The basic idea is to obtain solutions of (3.1) as smooth perturbations of composite orbits of the decoupled limiting equations (3.3) and (3.4). We make the following assumptions.

(i) The equation $g(x, y) = 0$ has a smooth manifold of solutions. Let $\mathcal{C}$ be a compact submanifold of this manifold which is given as a graph of a C^r function $h : U \subset \mathbf{R}^m \rightarrow \mathbf{R}^k$.

(ii) There exist integers k_s and k_u, with $k = k_s + k_u$, such that the matrix $\partial g(x, h(x))/\partial y$ has k_s eigenvalues with negative real part uniformly bounded away from zero, and k_u eigenvalues with positive real part uniformly bounded away from zero for all $x \in U$.

Interpret the following assertions in an appropriate neighbourhood V of $\mathcal{C}$. Under the above assumptions the reduced problem (3.3) defines a flow on $\mathcal{C}$. Additionally, $\mathcal{C}$ is an invariant manifold of fixed points for the flow defined by (3.4). We denote the k_s-dimensional local stable manifold of $q \in \mathcal{C}$ by $\mathcal{F}^s(q)$ and, similarly, the k_u-dimensional local unstable manifold by $\mathcal{F}^u(q)$. It is shown in [2, Thm. 9.1] that for sufficiently small ε the manifold $\mathcal{C}$ perturbs to a locally invariant center-like manifold $\mathcal{C}_\varepsilon$ with a $m+k_s$-dimensional center-stable manifold $\mathcal{C}^s_\varepsilon$ and a $m+k_u$-dimensional center-unstable manifold $\mathcal{C}^u_\varepsilon$. The flow on $\mathcal{C}_\varepsilon$ is a regular perturbation of the reduced problem on $\mathcal{C}$. Furthermore, there exist invariant foliations of $\mathcal{C}^s_\varepsilon$ and $\mathcal{C}^u_\varepsilon$ by k_s-dimensional manifolds $\mathcal{F}^s_\varepsilon(q)$ and k_u-dimensional manifolds $\mathcal{F}^u_\varepsilon(q)$, $q \in \mathcal{C}_\varepsilon$, respectively. The dependence of these manifolds on ε is C^{r-1}, even at $\varepsilon = 0$. For details we refer to Fenichel [2]. Background material can be found in Guckenheimer and Holmes [6] and Hirsch, Pugh, and Shub [8].

In our analysis we encounter for a given flow compact invariant manifolds $\mathcal{M}$ with boundary $\partial\mathcal{M}$. In order to apply results from invariant manifold theory $\mathcal{M}$ has to be overflowing (inflowing) invariant. In the following we do not mention this condition explicitly each time, but always assume that it has been achieved by a standard local modification of the flow near $\partial\mathcal{M}$.

Let $p \in \mathcal{C}$ be a hyperbolic fixed point of (3.3). Let $\Gamma^s(p)$ and $\Gamma^u(p)$ denote the local stable and unstable manifold of p for the reduced problem. We define the singular stable and unstable manifold, respectively, by

$$W^s(p) = \bigcup_{q \in \Gamma^s(p)} \mathcal{F}^s(q), \qquad W^u(p) = \bigcup_{q \in \Gamma^u(p)} \mathcal{F}^u(q). \tag{3.5}$$

It follows from [2, Thm. 12.2] that $W^s(p)$ perturbs smoothly to the stable manifold $W^s_\varepsilon(p)$ of the hyperbolic fixed point p of (3.1) for small ε. Similarly, $W^u(p)$ perturbs smoothly to the unstable manifold $W^u_\varepsilon(p)$.

Usually, the manifold of solutions of $g(x, y) = 0$ has several branches. Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be two manifolds, which satisfy conditions (i) and (ii). Let $p_1 \in \mathcal{C}_1$ be a hyperbolic fixed point of (3.3) with unstable manifold Γ^u_1, and $p_2 \in \mathcal{C}_2$ be a hyperbolic fixed point of (3.3) with stable manifold Γ^s_2. Assume that $q_1 \in \Gamma^u_1$ and $q_2 \in \Gamma^s_2$ are connected by a heteroclinic orbit η of (3.4). Let $\gamma_1 \subset \Gamma^u_1$ be the solution segment of (3.3) connecting p_1 and q_1, similarly, $\gamma_2 \subset \Gamma^s_2$ connects q_2 and p_2. Then, we call $\omega = \gamma_1 \cup \eta \cup \gamma_2$ the singular orbit connecting p_1 and p_2. Note, that we allow $p_1 = q_1$ or $p_2 = q_2$. Clearly, $\eta \subset W^u(p_1) \cap W^s(p_2)$ holds.

THEOREM 3.1. *Under the assumptions made in this section, assume that the singular unstable manifold $W^u(p_1)$ and the singular stable manifold $W^s(p_2)$ intersect transversally and locally unique along the orbit ω. Then there exists a locally unique transversal heteroclinic orbit ω_ε of* (3.1) *for small ε. The orbit ω_ε has a transition layer and is uniformly close to ω.*

The theorem follows from the smooth dependence of the involved manifolds on ε, and from the stability of transversal intersection under small perturbations. If the singular stable and unstable manifolds intersect transversally in an s-dimensional manifold $\mathcal{D}$ of singular heteroclinic orbits, we obtain the existence of a locally unique s-dimensional manifold $\mathcal{D}_\varepsilon$ of heteroclinic orbits of (3.1) for small ε. Details and examples are given in Szmolyan [13]. For suitable parametrizations ω_ε converges uniformly to ω in C^1-norm, away from the location of the transition layer.

LEMMA 3.2. *Under the above assumptions we parametrize γ_1 and γ_2 by $t \in \mathbf{R}_0^-$ and $t \in \mathbf{R}_0^+$, respectively, such that $\gamma_1(0) = q_1$ and $\gamma_2(0) = q_2$ hold. Let $\mathcal{V}$ be a manifold which intersects η transversally. We parametrize ω_ε such that $\omega_\varepsilon = \omega_\varepsilon(t)$, $t \in \mathbf{R}$, $\omega_\varepsilon(0) = \omega_\varepsilon \cap \mathcal{V}$ holds. Then*

$$\lim_{\varepsilon \to 0} \omega_\varepsilon(t) = \begin{cases} \gamma_1(t), & t < 0, \\ \gamma_2(t), & t > 0, \end{cases}$$

holds. The convergence is uniform in $C^1((-\infty, -\delta])$ and $C^1([\delta, \infty))$ for $\delta > 0$.

The proof of this intuitive result, based on the estimates in Theorem 9.1, (iii), [2], is straightforward and is omitted. We will also encounter the following simpler situation, which is contained in Theorem 3.1, if $\mathcal{C}_1 = \mathcal{C}_2$ holds.

COROLLARY 3.3. *Assume that there exist two hyperbolic fixed points p_1 and p_2 in a manifold $\mathcal{C}$, which satisfies conditions* (i), (ii). *Suppose that p_1 and p_2 are connected by a transversal heteroclinic orbit ω of* (3.3). *Then there exists a transversal heteroclinic orbit $\omega_\varepsilon \subset \mathcal{C}_\varepsilon$ of* (3.1) *for small ε. The orbit ω_ε has no transition layer and is uniformly close to ω in $C^1(\mathbf{R})$-norm.*

In higher dimensions it is usually difficult to prove transversality of the intersection of the singular stable and unstable manifolds. An analytic method to prove transversality based on an application of the Melnikov integral is given in [13]. However, we shall see that for detonation and deflagration waves the transversality condition is trivially satisfied. In §5 we shall need some extensions of the above results because (1.6) has a one-dimensional manifold of nonhyperbolic, fixed points due to the ignition temperature assumption.

4. Structure of gasdynamic shocks.

The Rankine–Hugoniot conditions for gasdynamic shocks for an ideal gas are derived from the three conservation laws in (1.3). Viscous profiles for a gasdynamic shock are heteroclinic orbits of the first two equations in (1.6). In our scaling the equations are

$$\begin{aligned} \mu u_x &= u - 1 + \frac{1}{\gamma M^{\star 2}} \left(\frac{T}{u} - 1 \right), \\ \lambda T_x &= T - 1 - \frac{\gamma - 1}{\gamma}(T - u) + qZ - \frac{(\gamma - 1)M^{\star 2}}{2}(u - 1)^2. \end{aligned} \tag{4.1}$$

Note, that the constant qZ vanishes in a nonreactive gas; however, we include this term for later convenience. For $Z = 0$, (4.1) are the equations for a viscous profile connecting the states corresponding to the unscaled states $(u_\star, T_\star)$ and $(u^\star, T^\star)$. We write (4.1) as

$$\begin{aligned} \mu \dot{u} &= f(u, T), \\ \lambda \dot{T} &= g(u, T), \end{aligned} \tag{4.2}$$

where the superscript "$\cdot$" denotes differentiation with respect to x. The equations $f(u, T) = 0$ and $g(u, T) = 0$ describe two parabolas in the u, T-plane, which we denote by F and G, respectively. The strict inequality (2.8) implies that the two parabolas intersect as shown in Fig. 2 for $Z \geq 0$. For increasing values of Z, the minimum of the parabola G decreases. More specifically, the following properties are easily verified for $u > 0$ and $T > 0$:

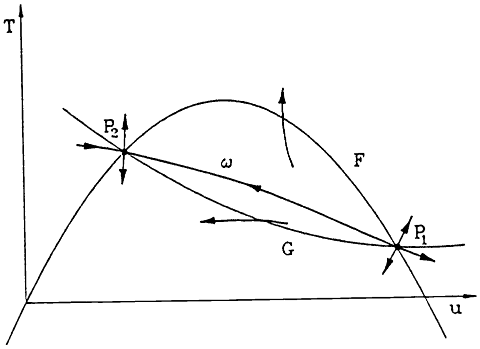


FIG. 2. *Shock structure.*

(a) $f_T > 0$, $g_T > 0$;

(b) there exist exactly two fixed points $P_1 = (u_1, T_1)$ and $P_2 = (u_2, T_2)$ with $u_1 > u_2$;

(c) $g_u > 0$ on G for $u_2 \leq u \leq u_1$;

(d) $g_u/g_T < f_u/f_T$ at P_1, and $g_u/g_T > f_u/f_T$ at P_2;

(e) there exists $u_2 < u_0 < u_1$, such that $f_u < 0$ for $u < u_0$, $f_u > 0$ for $u > u_0$ hold on F.

For sufficiently large Z, the values of T_1 become negative, which makes the problem physically meaningless. Properties (a)–(d) imply that P_1 is an unstable node and that P_2 is a saddle. Furthermore, all orbits cross the boundary of the region between F and G in the outward direction. Since the stable manifold of P_2 points into this region, we obtain the following lemma.

LEMMA 4.1. *In the above situation, there exists a transversal heteroclinic orbit ω connecting P_1 and P_2 for all values of μ and λ. The variables T and u are monotone along ω.*

This result is proved in Gilbarg [7] for general thermodynamics. If $\lambda, \mu \to 0$ simultaneously, the viscous profile converges to the shock wave. Gilbarg also analysed the limiting behavior of these viscous profiles as the ratios $\lambda/\mu \to 0$ and $\mu/\lambda \to 0$, respectively. We show how these limits fit into the framework developed in §3. In the first case we rewrite (4.2) as

$$\begin{aligned} \dot{u} &= f(u,T), \\ \varepsilon\dot{T} &= g(u,T), \end{aligned} \tag{4.3}$$

where $\varepsilon = \lambda/\mu$ is small. Thus, (4.3) is of the form (3.1); property (a) implies that conditions (i) and (ii) from §3 are satisfied on compact segments $\mathcal{C} \subset G$. The corresponding reduced problem is one-dimensional and has a (singular) heteroclinic orbit ω connecting P_1 and P_2 as shown in Fig. 3. In our graphics orbits of the reduced problem are indicated by single arrrows, orbits of the layer problem are indicated by double arrows. The transversality condition of Corollary 3.3 is trivially satisfied, and we conclude the existence of a transversal heteroclinic orbit ω_ε of (4.3) uniformly close to ω for small ε. This proves uniform convergence of the orbits given by Lemma 4.1 to ω as the ratio $\lambda/\mu \to 0$.

In the second case we rewrite (4.2) as

$$\varepsilon\dot{u} = f(u,T), \qquad \dot{T} = g(u,T), \tag{4.4}$$

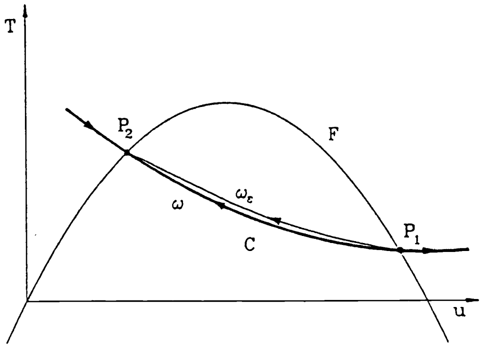


FIG. 3. $\lambda/\mu \to 0$.

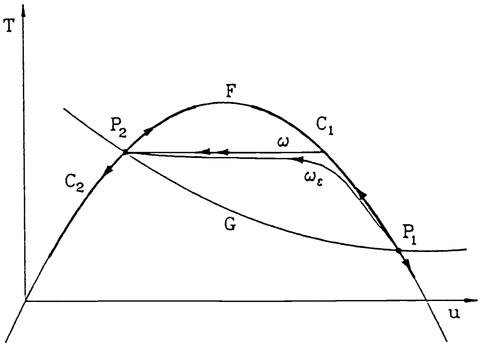


FIG. 4. $\mu/\lambda \to 0$.

where $\varepsilon = \mu/\lambda$ is small. In this case the manifold of the reduced problem is the parabola F, however, conditions (i), (ii) from §3 are only satisfied for u bounded away from u_0 due to property (e). Thus, the reduced problem is defined on compact segments $\mathcal{C}_1$ and $\mathcal{C}_2$ of the right and left branch of F (compare Fig. 4); P_1 and P_2 are both unstable fixed points of the one-dimensional reduced problem on $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively. It is easy to see, that there exists a singular heteroclinic orbit ω, because each point in $\mathcal{C}_1$ is connected to a point $\mathcal{C}_2$ by a (horizontal) heteroclinic orbit of the corresponding layer problem (3.4). The transversality condition of Theorem 3.1 is trivially satisfied, and we conclude the existence of a transversal heteroclinic orbit ω_ε of (4.4) uniformly close to ω for small ε. This proves uniform convergence of the orbits given by Lemma 4.1 to ω as the ratio $\mu/\lambda \to 0$.

5. Detonation and deflagration waves. In this section we consider (2.9)–(2.12) for small μ, λ and D with fixed ratios Pr and Le, i.e., we set $\mu = \varepsilon\hat{\mu}$, $\lambda = \varepsilon\hat{\lambda}$ and $D = \varepsilon\hat{D}$, where ε is small; $\hat{\mu}$, $\hat{\lambda}$, and $\hat{D}$ are positive constants. We assume that $\phi(T)$ is smooth, with ignition temperature T_i, i.e., $\phi(T) = 0$ for $T \leq T_i$. Thus, we obtain

$$\varepsilon\hat{\mu}u_x = u - 1 + \frac{1}{\gamma M^{\star 2}}\left(\frac{T}{u} - 1\right), \tag{5.1}$$

$$\varepsilon\hat{\lambda}T_x = T - 1 - \frac{\gamma - 1}{\gamma}(T - u) + qZ - \frac{(\gamma - 1)M^{\star 2}}{2}(u - 1)^2, \tag{5.2}$$

$$\varepsilon \hat{D} Y_x = u(Y - Z), \tag{5.3}$$

$$Z_x = -\frac{Y}{u}\phi(T). \tag{5.4}$$

Obviously, (5.1)–(5.4) is of the form (3.1), with the slow variable Z, and the fast variables u, T, and Y. Therefore, the methods from §3 are applicable. Mathematically, the equations define a smooth dynamical system on the phase space $\mathbf{R}^+ \times \mathbf{R} \times \mathbf{R} \times \mathbf{R}$, where we keep in mind that only $T > 0$ and $Y \in [0,1]$ have physical meaning. By setting $\varepsilon = 0$, we obtain the reduced problem

$$0 = u - 1 + \frac{1}{\gamma M^{\star 2}}\left(\frac{T}{u} - 1\right), \tag{5.5}$$

$$0 = T - 1 - \frac{\gamma - 1}{\gamma}(T - u) + qZ - \frac{(\gamma - 1)M^{\star 2}}{2}(u - 1)^2, \tag{5.6}$$

$$0 = u(Y - Z), \tag{5.7}$$

$$Z_x = -\frac{Y}{u}\phi(T). \tag{5.8}$$

Equations (5.5)–(5.7) define the one-dimensional manifold $\mathcal{C}$, on which (5.8) defines a flow. Note that (5.5) does not depend on Y, Z, and (5.6) does not depend on Y. For $u > 0$, (5.7) implies $Y = Z$. Thus, we can visualize $\mathcal{C}$ in three-dimensional u, T, Z-space. In the u, T-plane (5.5) describes the parabola F from §4, in u, T, Z-space (5.5) is a cylinder $\mathcal{P}$ with crossection F for each $Z \in \mathbf{R}$. Note that the parabola F, and hence $\mathcal{P}$, correspond to the Raleigh line in Fig. 1. By using (5.5) to eliminate u^2 in (5.6) we obtain

$$0 = T\frac{\gamma + 1}{2\gamma} + qZ + u\frac{\gamma - 1}{2\gamma}(1 + \gamma M^{\star 2}\ ) - 1 - \frac{(\gamma - 1)M^{\star 2}}{2}. \tag{5.9}$$

This equation describes a plane $\mathcal{K}$, on which the values of T decrease for increasing u and Z. Thus, $\mathcal{P}$ and $\mathcal{K}$ intersect in the parabola $\mathcal{C}$ as shown in Fig. 5. Clearly, all fixed points of (5.1)–(5.4) must lie on $\mathcal{C}$. Due to (2.8) the Z coordinate of the vertex of $\mathcal{C}$ lies at $Z = Z_0 < 0$; for a Chapman–Jouget process $Z_0 = 0$ holds. We choose compact segments $\mathcal{C}_1$ and $\mathcal{C}_2$ on each branch of the smooth one-dimensional manifold $\mathcal{C}$, which can be parametrized by Z away from the vertex, i.e.,

$$\mathcal{C}_k = \{(u_k(Z), T_k(Z), Z, Z)\ :\ Z \in I\}, \qquad k = 1, 2.$$

The closed interval I has to be chosen, such that $u_2(Z)$ is bounded away from zero. The layer equations with the fast independent variable $\xi = x/\varepsilon$ are

$$\hat{\mu} u_\xi = u - 1 + \frac{1}{\gamma M^{\star 2}}\left(\frac{T}{u} - 1\right), \tag{5.10}$$

$$\hat{\lambda} T_\xi = T - 1 - \frac{\gamma - 1}{\gamma}(T - u) + qZ - \frac{(\gamma - 1)M^{\star 2}}{2}(u - 1)^2, \tag{5.11}$$

$$\hat{D} Y_\xi = u(Y - Z), \tag{5.12}$$

$$Z_\xi = 0. \tag{5.13}$$

The layer problem is the gasdynamic shock problem (4.1) coupled to the trivial equation (5.12), which adds one unstable dimension for $u > 0$. The discussion in §4 implies

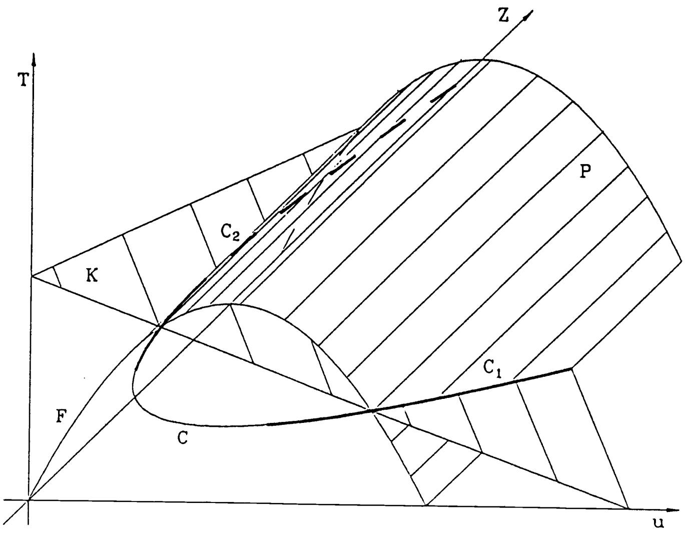


FIG. 5. *Phase space.*

that all points $P_1 \in \mathcal{C}_1$ are unstable nodes of (5.10)–(5.12) with three-dimensional unstable manifolds $\mathcal{F}^u(P_1)$. All points $P_2 \in \mathcal{C}_2$ are saddle points of (5.10)–(5.12) with one-dimensional stable manifolds $\mathcal{F}^s(P_2)$ and two-dimensional unstable manifolds $\mathcal{F}^u(P_2)$. For $Z \in I$ the real parts of the corresponding eigenvalues of the linearization are uniformly bounded away from zero. Thus, conditions (i) and (ii) from §3 are satisfied on $\mathcal{C}_1$ and $\mathcal{C}_2$, and the invariant manifold theory from §3 is applicable.

Lemma 3.2 implies that for $Z \in I$ each point $P_1(Z) \in \mathcal{C}_1$ is connected to the point $P_2(Z) \in \mathcal{C}_2$ by a heteroclinic orbit $\omega(Z)$ of (5.10)–(5.13). The u, T-component of the orbit $\omega(Z)$ is the viscous profile for the gasdynamic shock connecting $(u_1(Z), T_1(Z))$ and $(u_2(Z), T_2(Z))$, the coordinate $Y = Z$ is constant. Thus, there exists a smooth, two-dimensional manifold $\mathcal{S}$ of orbits connecting $\mathcal{C}_1$ and $\mathcal{C}_2$ (see Fig. 6).

The next step in our construction of singular detonation and deflagration waves is the analysis of the reduced problem (5.5)–(5.8), which is governed by the one-dimensional equation

$$(5.14) \qquad Z_x = -\frac{Z}{u_k(Z)}\,\phi(T_k(Z)), \qquad k = 1, 2.$$

Fixed points of (5.14) have to satisfy $Z = 0$ or $\phi(T_k(Z)) = 0$, $k = 1, 2$. Clearly, $Z = 0$ gives the fixed points $P_\star \in \mathcal{C}_1$ and $P^\star \in \mathcal{C}_2$. The reaction rate $\phi(T_k(Z)) = 0$, whenever $T_k(Z) \leq T_i$, $k = 1, 2$, holds.

On $\mathcal{C}_1$ the temperature is a strictly decreasing function of Z; therefore, there exists at most one solution Z_1 of the equation $T_1(Z) = T_i$, and $\phi(T_1(Z)) = 0$ for $Z \geq Z_1$. In this section we assume the existence of such $Z_1 > 0$. The case that $\mathcal{C}_1$ consists entirely of fixed points of (5.14) is less interesting and can be analysed similarly. On $\mathcal{C}_2$ the temperature is a strictly decreasing function of Z if and only if the point $P^\star$ lies on the left side of the vertex of the parabola F in the u, T-plane; this holds for $\gamma M^{\star 2} < 1$. In the case $\gamma M^{\star 2} > 1$ the temperature on $\mathcal{C}_2$ increases to the maximum value $(1 + \gamma M^{\star 2})^2/(4\gamma M^{\star 2})$ at the top of $\mathcal{P}$ and decreases for larger values of Z. Therefore, the equation $T_2(Z) = T_i$ has either one or two solutions. In this section we discuss the case of a unique solution $Z_2 > 0$, the other case of

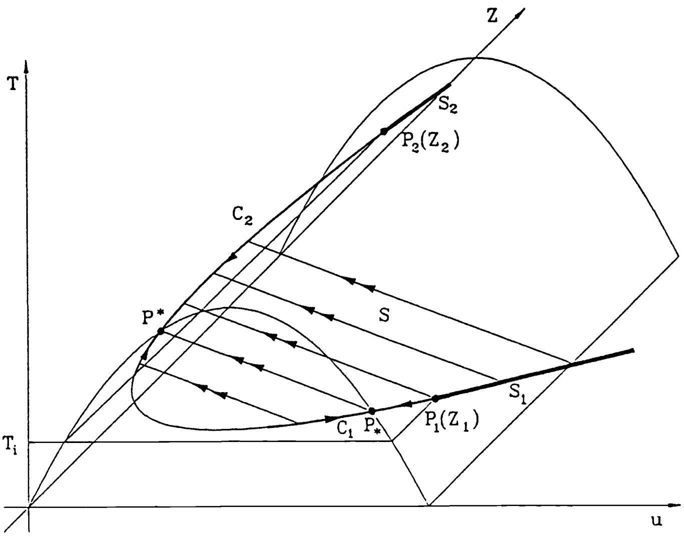


FIG. 6. *Singular invariant manifolds.*

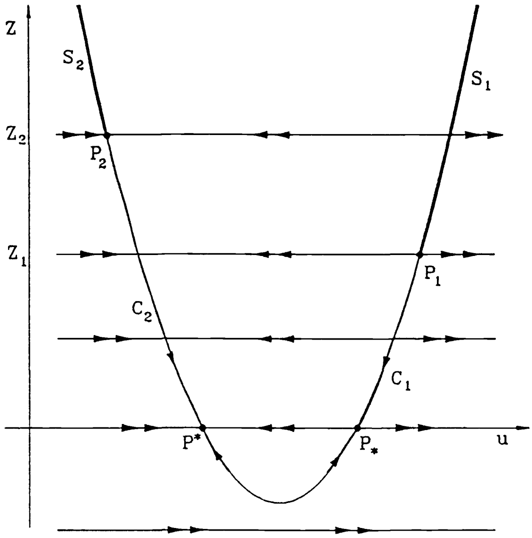


FIG. 7. *Singular orbits.*

incompletely burned states is discussed in §6. Since $T_2(Z) > T_1(Z)$ holds, we conclude $Z_2 > Z_1$. In this discussion we have implicitly assumed that the interval I has been choosen large enough such that the interval $[0, Z_2]$ is contained in the interior of I. Thus, we have the situation shown in Fig. 7. On C_1 (5.14) has the hyperbolic attracting fixed point $P_\star$ and a continuum of (necessarily nonhyperbolic) fixed points $S_1 = \{P_1(Z) : Z \in I, Z \geq Z_1\}$. Since there are no other fixed points on C_1, the point $P_1(Z_1)$ is connected to $P_\star$ by a heteroclinic orbit of (5.5)–(5.8). The situation on C_2 is similar. There exists a continuum of fixed points $S_2 = \{P_2(Z) : Z \in I, Z \geq Z_2\}$, and the point $P_2(Z_2)$ is connected to the attracting hyperbolic fixed point $P^\star$ by a heteroclinic orbit of (5.5)–(5.8).

The given unburned state P_- lies in S_1 for $M_-^2 > 1$, and in S_2 for $M_-^2 < 1$. However, it is clear that P_- is not a particularly distinguished fixed point of (5.1)–(5.4). Thus, we prefer to think of $P^\star$ as a given subsonic burned state, which determines $P_\star$, S_1, and S_2. Then, all points in S_1 are possible left states for detonation waves with burned states $P_\star$ or $P^\star$; all points in S_2 are possible left states for deflagration waves with burned states $P_\star$ or $P^\star$. We summarize our discussion in the following theorem.

THEOREM 5.1. *Under the assumptions made in this section, the following singular heteroclinic orbits connecting fixed points of* (5.1)–(5.4) *exist.*

(i) *The point $P_\star$ is connected to $P^\star$ by the heteroclinic orbit of* (5.10)–(5.13).

(ii) *The point $P_1(Z_1)$ is connected to $P_\star$ by the heteroclinic orbit of* (5.5)–(5.8).

(iii) *There exists a two-dimensional manifold $\mathcal{D}$ (with boundary) of singular orbits from $P_1(Z_1)$ to $P^\star$. These orbits consist of the solution segment of* (5.5)–(5.8) *from $P_1(Z_1)$ to $P_1(Z)$, the heteroclinic orbit of* (5.10)–(5.13) *from $P_1(Z)$ to $P_2(Z)$, and the solution segment of* (5.5)–(5.8) *from $P_2(Z)$ to $P^\star$ for $Z \in (0, Z_1]$. The boundary of $\mathcal{D}$ is formed by the singular orbit corresponding to $Z = Z_1$ and by the singular orbits from* (i) *and* (ii).

(iv) *For $Z_1 < Z < Z_2$ the point $P_1(Z)$ is connected to $P^\star$ by a heteroclinic orbit of* (5.10)–(5.13) *from $P_1(Z)$ to $P_2(Z)$ followed by the solution segment of* (5.5)–(5.8) *from $P_2(Z)$ to $P^\star$.*

(v) *For $Z \geq Z_2$ the point $P_1(Z)$ is connected to $P_2(Z)$ by the heteroclinic orbit of* (5.10)–(5.13).

(vi) *The point $P_2(Z_2)$ is connected to $P^\star$ by the heteroclinic orbit of* (5.5)–(5.8). *All these singular orbits lie in the manifold $\mathcal{S}$. There are no other singular heteroclinic orbits.*

The situation is shown in Fig. 7. In cases (iv) and (v) it is possible, that $T_1(Z)$ is negative; then the singular orbit is physically meaningless and only included for mathematical completeness. If we replace the heteroclinic orbits of the layer problem (5.10)–(5.13) in Theorem 5.1 by jump discontinuities in the fast gasdynamic variables u and T, then the theorem describes all solutions of the ZND-model (1.3) with right states $P^\star$ and $P_\star$. However, Fig. 7 reveals more details than the classical Chapman–Jouget diagram in Fig. 1. Furthermore, as should be obvious by now, Theorem 5.1 is the basis for the proof of the main theorem from §1, which follows, if we prove that the singular heteroclinic orbits perturb to heteroclinic orbits of (5.1)–(5.4) for small ε. Theorem 3.1 does not apply directly because all fixed points of (5.5)–(5.8) in S_1 and S_2 are nonhyperbolic. Clearly, this nongeneric situation, that could be destroyed by small perturbations, is caused by the ignition temperature assumption.

THEOREM 5.2. *Under the assumptions made in this section, there exists $\varepsilon_0 = \varepsilon_0(\hat{\mu}, \hat{\lambda}, \hat{D}) > 0$, such that the following assertions hold for* (5.1)–(5.4) *for $0 < \varepsilon < \varepsilon_0$.*

(i) *$P^\star$ and $P_\star$ are hyperbolic fixed points, with two-dimensional and one-dimensional stable manifolds, respectively. S_1 and S_2 are one-dimensional manifolds of fixed points.*

(ii) *The singular heteroclinic orbits from Theorem* 5.1, (i)–(vi) *perturb smoothly to heteroclinic orbits of* (5.1)–(5.4) *connecting the corresponding fixed points.*

(iii) *All connecting orbits lie in a smooth, two-dimensional, invariant manifold $\mathcal{S}_\varepsilon$, which is a graph over the u, Z-plane. The flow in $\mathcal{S}_\varepsilon$ is as shown in Fig.* 8.

Proof. The manifolds $\mathcal{C}_1$ and $\mathcal{C}_2$ satisfy conditions (i) and (ii) from §3, and we conclude the existence of the center-like manifolds $\mathcal{C}_{i,\varepsilon}$, their stable and unstable manifolds $\mathcal{C}^s_{i,\varepsilon}$ and $\mathcal{C}^u_{i,\varepsilon}$, and their invariant foliations $\mathcal{F}^s_{i,\varepsilon}$ and $\mathcal{F}^u_{i,\varepsilon}$, respectively, for

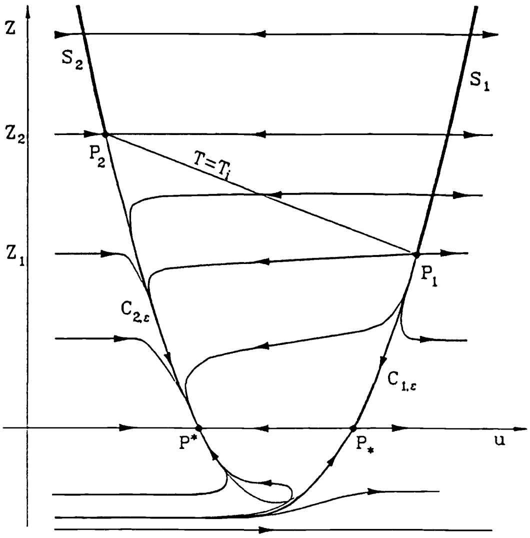


FIG. 8. *Flow in* $\mathcal{S}_\varepsilon$.

$i = 1, 2$, and ε sufficiently small. $\mathcal{C}^u_{1,\varepsilon}$ and $\mathcal{C}^s_{2,\varepsilon}$ are smooth perturbations of

$$\mathcal{C}^u_1 = \bigcup_{P_1 \in \mathcal{C}_1} \mathcal{F}^u(P_1), \qquad \mathcal{C}^s_2 = \bigcup_{P_2 \in \mathcal{C}_2} \mathcal{F}^s(P_2).$$

We conclude from Theorem 5.1 that the four-dimensional manifold $\mathcal{C}^u_1$ and the two-dimensional manifold $\mathcal{C}^s_2$ intersect transversally in the two-dimensional manifold $\mathcal{S}$ (see Fig. 6). The stability of transversal intersection proves the existence a smooth, invariant manifold

$$\mathcal{S}_\varepsilon = \mathcal{C}^u_{1,\varepsilon} \cap \mathcal{C}^s_{2,\varepsilon},$$

C^1-close to $\mathcal{S}$ for small ε. The manifold $\mathcal{S}$ is a graph, i.e., $T = T(u, Z)$ and $Y = Z$ on $\mathcal{S}$; thus, $\mathcal{S}_\varepsilon$ is a graph over the u, Z-plane.

Next, we discuss the existence of heteroclinic orbits in $\mathcal{S}_\varepsilon$. Since $P_\star$ and $P^\star$ are stable hyperbolic fixed points of (5.5)–(5.8), they must lie in $\mathcal{C}_{1,\varepsilon}$ and $\mathcal{C}_{2,\varepsilon}$, respectively, and satisfy assertion (i), which can be also verified by direct computation for arbitrary ε. The existence of a (purely gasdynamic) heteroclinic orbit connecting $P_\star$ and $P^\star$ follows either from Theorem 3.1 or from Lemma 4.1 by restricting the flow to the invariant subspace $Y = Z = 0$. The last argument implies also the existence of (purely gasdynamic) heteroclinic orbits connecting $P_1(Z)$ and $P_2(Z)$ for $Z \geq Z_2$.

Since S_1 and S_2 are sets of fixed points for (5.1)–(5.4), they are invariant. Thus, Theorem 9.1 in [2] implies $S_1 \subset \mathcal{C}_{1,\varepsilon}$ and $S_2 \subset \mathcal{C}_{2,\varepsilon}$. Since there are no other fixed points, the one-dimensional flow on $\mathcal{C}_{1,\varepsilon}$ must connect $P_1(Z_1)$ and $P_\star$. Similarly, there exists an orbit from $P_2(Z_2)$ to $P^\star$ in $\mathcal{C}_{2,\varepsilon}$. This proves the persistence of the orbits given by (ii) and (vi) in Theorem 5.1. Thus, the region on $\mathcal{S}_\varepsilon$ bounded by the orbits from $P_\star$ to $P^\star$, from $P_1(Z_1)$ to $P_\star$, from $P_2(Z_2)$ to $P^\star$, from $P_1(Z_2)$ to $P_2(Z_2)$, and by the curve of fixed points $\{P_1(Z) : Z_1 \leq Z \leq Z_2\}$ is invariant (see Fig. 8). Since there are no other fixed points, all points in this region belong to the stable manifold of $P^\star$ and assertion (ii) follows. This proves that the flow is as shown in Fig. 8, which shows the flow in the union of $\mathcal{C}^s_{2,\varepsilon}$ and the smooth extension of $\mathcal{S}_\varepsilon$ in $\mathcal{C}^u_{1,\varepsilon}$; $\mathcal{S}_\varepsilon$ is the region between $\mathcal{C}_{1,\varepsilon}$ and $\mathcal{C}_{2,\varepsilon}$. □

For a given unburned state P_- the main result in §1 follows from Lemma 2.1, Theorem 5.2, and Lemma 3.2. Weak detonation waves and weak deflagration waves are heteroclinic orbits of the type described in Corollary 3.3; they have no internal layers. Strong detonation waves are orbits of the type described in Theorem 3.1; the gasdynamic shock corresponds to an internal layer.

6. Discussion and extensions. Our analysis proves the existence of weak detonation waves and weak deflagration waves for unburned temperatures exactly at ignition. The process is driven by the chemical reaction; the gasdynamic variables are close to equilibrium. Strong detonation waves with unburned temperature below ignition are unique; the temperature is raised above ignition by the gasdynamic shock. Then the reaction proceeds like a weak deflagration. There is a one-parameter family of strong detonation waves with unburned temperature exactly at ignition. In one of the two extreme cases the detonation wave has the usual structure of a gasdynamic shock followed by a weak deflagration. The other limiting case is a weak detonation followed by a nonreactive gasdynamic shock. In the intermediate cases the process starts like a weak detonation; then the temperature is raised by a gasdynamic shock, and the process is completed like a weak deflagration.

In the following we discuss the monotonicity properties of temperature and pressure. The discussion is based on the fact that strict monotonicity of a variable along a singular orbit is preserved for small ε. For $\gamma M^{\star 2} < 1$, temperature is strictly increasing along all singular orbits. Thus, temperature is strictly increasing along all detonation and deflagration waves for $\gamma M^{\star 2} < 1$ and small ε. For $\gamma M^{\star 2} > 1$, temperature is strictly increasing along weak detonations. However, on the deflagration branch temperature is increasing for temperatures close to ignition but decreasing near $P^\star$. Therefore, the temperature has a maximum in the interior of weak deflagration and strong detonation waves for $\gamma M^{\star 2} > 1$ and small ε. Pressure is strictly increasing along weak detonation waves and strictly decreasing along weak deflagration waves for small ε. Hence, for a strong detonation wave the pressure attains its maximum after the gasdynamic shock; this peak in pressure corresponds to the familiar Neumann spike [16].

In §5 we mentioned the possibility of incompletely burned states. For $\gamma M^{\star 2} > 1$ it is possible that the equation $T_2(Z) = T_i$ has two solutions $0 < Z_2' < Z_2$. In this case $T^\star < T_i$ holds, and, clearly, the temperature on $\mathcal{C}_1$ is below ignition. Thus, complete combustion is impossible. However, in this case a weak deflagration and strong detonations with a partial burned state exist for small ε. More precisely, there exists a weak deflagration wave with unburned state $P_2(Z_2)$ and burned state $P_2(Z_2')$. There exists a family of strong detonation waves with unburned state $P_1(Z)$ and burned state $P_2(Z_2')$ for $Z_2' < Z < Z_2$. The existence of these waves follows from the existence of the corresponding singular heteroclinic orbits similar to the proof of Theorems 5.2. The existence of weak deflagration waves with an incomplete burned state is discussed in Wagner [15]; the existence of strong detonation waves with a partial burned state seems to be new.

In our analysis we have assumed that $\hat{\mu}$, $\hat{\lambda}$, and $\hat{D}$ in (5.1)–(5.4) are arbitrary constants. Since the values of these parameters have no impact on the analysis of §5, our results are uniformly valid as long as $\hat{\mu}$, $\hat{\lambda}$, and $\hat{D}$ are bounded from below and from above by some positive constants. More precisely, for fixed values $0 < \alpha < \beta < \infty$ a constant $\varepsilon_0(\alpha, \beta) > 0$ exists, such that the assertions of Theorem 5.2 hold for

$0 < \varepsilon < \varepsilon_0$ for all $\alpha < \hat{\mu} < \beta$, $\alpha < \hat{\lambda} < \beta$, and $\alpha < \hat{D} < \beta$.

The case that $(\mu, \lambda, D) \to 0$ along a smooth curve in parameterspace with asymptotically bounded ratios can be treated in the following way. We set

$$\mu = \varepsilon\hat{\mu}(\varepsilon), \quad \lambda = \varepsilon\hat{\lambda}(\varepsilon), \quad D = \varepsilon\hat{D}(\varepsilon)$$

in (2.9)–(2.12), where $\hat{\mu}$, $\hat{\lambda}$, $\hat{D}$ are smooth, positive functions on an interval $[0, \varepsilon_1]$ with $\varepsilon_1 > 0$. This gives a system of the form (3.1), where the function g depends smoothly on ε. The theory outlined in §3 applies also to the case that the right-hand side in (3.1) depends smoothly on ε (see [2], [13]). The corresponding reduced problem is just (5.5)–(5.8), and the layer problem is (5.10)–(5.13) with $\hat{\mu} = \hat{\mu}(0)$, $\hat{\lambda} = \hat{\lambda}(0)$, and $\hat{D} = \hat{D}(0)$, i.e., we have exactly the situation analysed in §5. The particular simple dependence of the right-hand side g on ε, and the results in [13] imply that there exists $\varepsilon_0 > 0$, such that the assertions of the Main Theorem in §1 hold in this situation as well for all ε, $0 < \varepsilon < \varepsilon_0$.

The methods used in this paper can be applied to various other possible (less physical) singular limits, i.e., to cases where μ, λ, and D are of different orders of magnitude. In Gasser [5] a similar analysis in the case $\mu = \varepsilon\hat{\mu}$, $\lambda = \varepsilon\hat{\lambda}$, and $D = \hat{D}$ with positive constants $\hat{\mu}$, $\hat{\lambda}$, $\hat{D}$, and ε small is given. For $\varepsilon = 0$ this problem decouples into a two-dimensional reduced problem describing chemistry and diffusion, and a two-dimensional layer problem describing pure gasdynamics. By applying the methods of §3 we obtain essentially the same results on the existence of detonation and deflagration waves.

For strong shocks $\hat{\mu}$ and $\hat{\lambda}$ may depend on the gasdynamic variables. If we consider $\hat{\mu} = \hat{\mu}(u, T, Y)$, $\hat{\lambda} = \hat{\lambda}(u, T, Y)$, and $\hat{D} = \hat{D}(u, T, Y)$ to be smooth functions, which are bounded from below and from above by some positive constants, all our results remain valid, since these positive factors in (5.1)–(5.4) do not change the geometry and the analysis of the problem.

REFERENCES

[1] R. Courant and K.O. Friedrichs, *Supersonic Flow and Shock Waves*, Applied Mathematical Sciences, 21, Springer, New York, 1948.

[2] N. Fenichel, *Geometric singular pertubation theory*, J. Differential Equations, 31 (1979), pp. 53–98.

[3] W. Fickett and W. Davis, *Detonation*, University of California Press, Berkeley, Los Angeles, CA, 1979.

[4] R. A. Gardner, *On the detonation of a combustible gas*, Trans. Amer. Math. Soc., 277 (1983), pp. 431–468.

[5] I. Gasser, *Viskose Profile für Deflagrations—und Detonationswellen*, Thesis, TU-Wien, 1991.

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

[7] D. Gilbarg, *The existence and limit behavior of the one-dimensional shock layer*, Amer. J. Math., 73 (1951), pp. 256–274.

[8] M. W. Hirsch, C. C. Pugh, and M. Shub, *Invariant manifolds*, Lecture Notes in Math., 583, Springer-Verlag, New York, Berlin, 1979.

[9] J. O. Hirschfelder and C. C. Curtiss, *Theory of detonations.* I. *Irreversible unimolecular reaction*, J. Chem. Phys., 28 (1958), pp. 1130–1147.

[10] P. Holmes and D. S. Stewart, *The existence of one dimensional steady detonation waves in a simple model problem*, Stud. Appl. Math., 66 (1982), pp. 121–143.

[11] G. C. Lu and G. S. S. Ludford, *Asymptotic analysis of plane steady detonations*, SIAM J. Appl. Math., 42 (1982), pp. 625–635.

[12] J. Smoller, *Shock Waves and Reaction—Diffusion Equations*, Grundlehren Math. Wiss., 258, Springer, New York, Berlin, 1983.

[13] P. Szmolyan, *Transversal heteroclinic and homoclinic orbits in singular pertubation problems*, J. Differential Equations, 92 (1991), pp. 252–281.

[14] D. H. Wagner, *The existence and behavior of viscous structure for plane detonation waves*, SIAM J. Math. Anal., 20 (1989), pp. 1035–1054.

[15] ———, *Detonation waves and deflagration waves in the one dimensional* ZND-*model for high Mach number combustion*, IMA-preprint 498, Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, MN, 1989.

[16] F. A. Williams, *Combustion Theory*, Benjamin/Cummings, Menlo Park, CA, 1985.

[17] W. W. Wood, *Existence of detonations for small values of the rate parameter*, Phys. Fluids, 4 (1961), pp. 46–60.

SIAM J. MATH. ANAL.
Vol. 24, No. 4, pp. 987–1008, July 1993

TRAVELLING WAVES FOR MUTUALIST SPECIES*

KONSTANTIN MISCHAIKOW† AND VIVIAN HUTSON‡

Abstract. A reaction-diffusion model for n mutalistic species is considered. The existence of a travelling wave analogous to a bistable wave for a single species is proved. Stability and uniqueness of the wave is considered, and the question of the "dominance" of the equilibria is discussed.

Key words. travelling waves, mutualism, symbiosis, reaction-diffusion, Conley index, connected simple systems

AMS subject classifications. 35K57, 92D25

1. Introduction.

Travelling wave problems for two interacting species whose dynamics are governed by a pair of reaction-diffusion equations have been much studied recently. To quote two examples only, Dunbar [5] has considered this problem for a predator-prey model, and Gardner [11] for a pair of competing species. It appears that a third class of model, that of mutualistic (sometimes called symbiotic) interactions, has not been considered from this point of view. However, it has been argued (see Hutson [13] and Hutson, Law, and Lewis [14]) that perhaps somewhat in contrast with the situation for the above types of interaction it is particularly important in this case to include diffusion to avoid the somewhat paradoxical conclusions predicted by ordinary differential equation models of mutualism when the mutualism is obligate (that is neither species can exist on its own). Our first objective then is to show that travelling waves exist for obligate mutualists, indeed even for a system of n obligate mutualists, and to discuss some of the properties of these waves.

We consider a system of reaction-diffusion equations

$$\frac{\partial u_i}{\partial t} = \mu_i \frac{\partial^2 u_i}{\partial x^2} + h_i(u) \qquad (i = 1, \dots, n), \tag{1.1}$$

where $u = (u_1, \dots, u_n)$, with spatial domain $\mathbf{R}$, the μ_i (assumed > 0) being diffusion coefficients. The phase plane (when $n = 2$) for the reaction system is described in Fig. 1, the detailed assumptions concerning the h_i being given in assumption (H1)–(H5) in §2. The reaction phase plane suggests that there is a broad analogy with the one species bistable case treated, for example, in Fife [6, pp. 106–109], based on the equation

$$\frac{\partial u}{\partial t} = \mu \frac{\partial^2 u}{\partial x^2} + m(u),$$

where m has the zeros $0, u^*, 1$ with $0 < u^* < 1$, $m(u) < 0$ for $0 < u < u^*$, and $m(u) > 0$ for $u^* < u < 1$. We shall show that this is indeed the case, proving in §4 that for some velocity c and arbitrary n, there is a monotone travelling wave (which we call a bistable wave) from a stable critical point A to another stable critical point

*Received by the editors October 24, 1990; accepted for publication (in revised form) October 19, 1992.

†Center for Dynamical Systems and Nonlinear Studies, School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332. This author was partially supported by National Science Foundation grant #DMS 8901273.

‡Department of Applied Mathematics, The University of Sheffield, Sheffield S102UN, England.

B. In §5 we demonstrate that for c large enough there is a travelling wave from A to C, and that a monotone such wave exists in the case $n = 2$.

A variety of techniques have been used to prove the existence of travelling waves, ranging from topological methods (Conley and Gardner [4], Gardner [10], [11]) to shooting methods based on Wazewski's principle (Dunbar [5]). Our second objective is to further develop the first class of methods, with the aim of treating a considerably wider class of systems, while at the same time reducing the technicalities to a minimum. The approach here may be compared with the use of classical degree theory, where the technicalities on which the theory rests do not appear in the results (such as the Leray–Schauder fixed point theorem). The abstract results presented in the Appendix are based on a development of the Conley index theory, Conley's connection matrix, and transition matrices. We then deduce in §3 two abstract continuation theorems which allow us to prove our existence results by carrying out a homotopy of the system to a simpler one. It is necessary here to construct "isolating neighborhoods" in phase space (analogous to open sets in degree theory) and to show that no bounded solution of interest can be internally tangent to the boundary of the neighborhood; this is done in §4.

The techniques presented in this paper are not new in the sense that Conley promoted the use of his index in proving existence of connecting orbits in the early 1970s and numerous people have since successfully used his ideas. Within the context of the types of travelling wave problems discussed here the most notable applications can be found in the works of Conley [2], Conley and Gardner [4], and Gardner [11]. For the single species problem one can find our approach sketched out in [2] (in particular, see the example IV.2.6). That these ideas could be applied to two-species problems was first demonstrated in [4] (competing species) and [11] (predator-prey). What we claim to show in this paper is, that in fact, this is a very general technique and can be straightforwardly applied to problems involving n-species. It is our contention that this approach can also be applied to higher-dimensional problems of mixed type, i.e., systems involving mutualist, competitive, and predator-prey species simultaneously (see [19] and [20]). It should also be remarked that whereas [4] and [11] apply a connection index to find connecting orbits, we use the connected simple system associated with the Conley index to obtain existence.[1] Results of [17] seem to imply that these two approaches are theoretically equivalent. Nevertheless, it is the authors' opinion that the connected simple system is the more basic concept, and thus, more natural to work with it. Furthermore, the global bifurcation results of [17] are obtained in the context of connected simple systems, and hence, they can be applied directly to the results obtained by our approach. (For examples of how these theorems can be applied to find nonmonotone bistable waves, the reader is referred to [20]).

2. Preliminaries. Consider the system of equations (1.1) and let $h_i(u) = u_i f_i(u)$. Take R to be the open rectangle in $\mathbf{R}^n_+$ with opposite vertices $A = (0, \ldots, 0)$ and $B = (1, \ldots, 1)$. We make the following hypothesis concerning the reaction system:

$$u_i' = h_i(u) = u_i f_i(u), \qquad (i = 1, \ldots, n) \tag{2.1}$$

corresponding to (1.1). It is assumed that the following assumptions hold in a neighborhood of $\bar{R}$ where all indices run over $1, \ldots, n$.

[1] Gardner has informed us that Conley originally considered using connected simple systems in their work, but then later formulated the connection index approach which first appeared in their joint paper.

(H1) $f_i \in C^2(\bar{R}, \mathbf{R})$.

(H2) $\partial f_i/\partial u_j > 0\,(i \neq j)$, that is, the interactions are mutualistic, and $\partial f_i/\partial u_i < 0$, that is, intraspecific competition holds.

(H3) A is a global attractor for the flow in any "face" $u_i = 0$, that is, the mutualism is obligate.

(H4) There are exactly three equilibria $A = (0,\dots,0)$, $B = (1,\dots,1)$ and $C = (c_1,\dots,c_n)$ with $0 < c_i < 1$ for all i. All three critical points are hyperbolic; A and B are stable and C has a one-dimensional unstable manifold.

(H5) The region $R_1 = \{u : f_i(u) > 0 \text{ for all } i = 1,\dots,n\}$ is connected and $B, C \in \bar{R}_1$. Also $R_2 = \{u : f_i(u) < 0 \text{ for all } i = 1,\dots,n\}$ is connected and $A, C \in \bar{R}_2$. Finally, $R_3 = \{u | u_i > 1, f_i(u) < 0 \text{ for all } i = 1,\dots,n\}$ is connected and $B \in \bar{R}_3$.

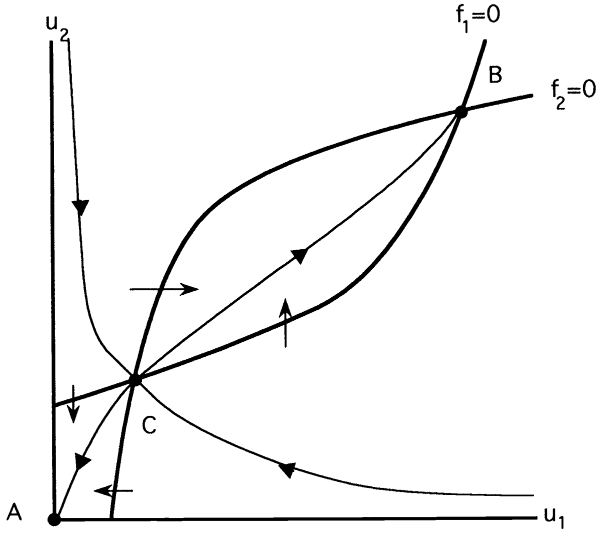


FIG. 1. *The phase plane for the reaction system when $n = 2$. The arrows give the vector field. The stable and unstable manifolds of C are shown with lighter lines.*

These assumptions yield simply the n-dimensional equivalent of the two-species mutualist system with phase plane as in Fig. 1. The following explicit example may help in visualising the type of model considered. For simplicity it is a three species case with some symmetry, but there is no essential difficulty in giving a similar n-species version or relaxing the symmetry. For any $\beta \in (\frac{1}{2}, 4)$ take

$$f_i(u) = -1 - u_i + (1 + 2\beta)\frac{\sum(i)}{1 + \beta\sum(i)},$$

where

$$\sum(i) = \sum_{j \neq i, j=1}^{3} u_j.$$

It is straightforward to show that there are exactly three equilibria A, C, B in $\mathbf{R}^3_+$ with coordinates $(0,0,0)$, $1/(2\beta)(1,1,1)$, and $(1,1,1)$, respectively. It is clear that on the diagonal $u_1 = u_2 = u_3$, starting from the origin, the sign of each f_i is negative, then changes to positive at C and then changes again to negative at B. On the other hand, consider the two-dimensional subsystem in a "face" $u_3 = 0$, for example. Then one can check easily that for β in the range prescribed above, the only equilibrium is the origin A, and drawing the null isoclines one can readily convince oneself that A is a global attractor. It follows rather easily that (H1)–(H5) hold.

There are several observations worth making at this point, the relevance of which will appear later in the proofs.

Remark 2.1. If $u \in B_\epsilon(B)$ such that $f_i(u) > 0$ for all $i = 1, \dots, n$, then $u_i < 1$. This follows from (H2) which determines the direction of ∇f_i and then (H5) which implies that $u_i < 1$ rather than $u_i > 1$.

Remark 2.2. Since the reaction system is a cooperative system it is a monotone system, and hence, using [12] by (H5) if u lies in the stable manifold of C, then it is impossible for $u_i - c_i > 0$ for all $i = 1, \dots, n$.

Remark 2.3. There exist connecting orbits from C to A and C to B which lie in the regions R_1 and R_2. To see why this is so, notice that a branch of the local unstable manifold of C lies in R_2. Assume for the moment that R_2 is positively invariant under the reaction flow. Since $u \in R_2$ implies $f_i(u) < 0$ for all $i = 1, \dots, n$, $u_i(t)$ is decreasing. However, the faces defined by $u_i = 0$ are invariant, hence $u_i(t) > 0$ for all $t > 0$. Since A is the only other critical point in $\bar{R}_2$, any point on this branch of the local unstable manifold of C limits to A. Thus, it only needs to be shown that $\bar{R}_2$ is positively invariant. Let $u \in \partial \bar{R}_2 \backslash \{C\}$. If $u_i = 0$ for some i then $\lim_{t \to \infty} u(t) = A$ by (H3). So assume $u_i > 0$ for all i, in which case $u \in \partial \bar{R}_2 \backslash \{C\}$ implies that there exists $j \in \{1, \dots, n\}$ such that $f_j(u) = 0$. On the other hand, since $u \neq C$, there exists f_i such that $f_i(u) \neq 0$. Of course, in this case $f_i(u) < 0$. Thus the tangent vector to the flow at u is given by $u_i' \leq 0$ and $u_j' = 0$. Note now that $u' \cdot \nabla f_j(u) < 0$, i.e., $f_j(u(\epsilon)) < 0$ for $0 < \epsilon \ll 1$, and hence the point u is entering the region R_2. A similar argument holds for the existence of a C to B orbit.

We seek travelling wave solutions of (1.1), that is solutions of the form $(u_1(x - ct), \dots, u_n(x - ct))$ where $c \in \mathbf{R}$ is the wave speed. Substitution into (1.1) yields the system of equations

$$\mu_i u_i'' + c u_i' + h_i(u) = 0 \qquad (i = 1, \dots, n), \tag{2.2}$$

where $u = (u_1, \dots, u_n)$, and "dash" denotes differentiation with respect to the argument $x - ct$ which, in an abuse of notation, we continue to refer to as t to fit the standard dynamical systems terminology. The corresponding first-order system is

$$u_i' = p_i \qquad (i = 1, \dots, n), \tag{2.3a}$$

$$\mu_i p_i' = -c p_i - h_i \qquad (i = 1, \dots, n). \tag{2.3b}$$

Notice that if $A = (a_1, \dots, a_n)$ is a critical point of the reaction system, then $A_0 = (A, 0) = (a_1, \dots, a_n, 0, \dots, 0)$ is a critical point of (2.3).

Following the discussion in the introduction, we consider in particular two types of travelling waves, which we describe as a Fisher wave and a bistable wave to emphasize the relation to the classical one-species types of wave; see [1] for basic facts concerning these waves, including existence proofs. A "connection" of two equilibria of the system (2.3) is a heteroclinic orbit from one critical point to the other. It is said to be monotone (here taken arbitrarily to be increasing always) if each $u_i(\cdot)$ is increasing. A Fisher wave is a monotone connection of A_0 to C_0, and a bistable wave is a connection of A_0 to B_0, where A and B are stable critical points for the reaction system.

Finally, it is convenient to note the following formulae derived from (2.3):

$$\mu_i \frac{dp_i}{du_i} = -c - p_i^{-1} h_i(u), \tag{2.4}$$

$$\frac{1}{2}\mu_i[p_i^2(u_{i1}) - p_i^2(u_{i0})] = -c\int_{u_{i0}}^{u_{i1}} p_i(s)ds - \int_{u_{i0}}^{u_{i1}} H_i(s)ds, \tag{2.5}$$

which hold along a monotone section of an orbit which is parameterized by u_i, where H_i denotes h_i in an obvious sense.

3. Two abstract continuation theorems. In this section we present a pair of abstract theorems upon which the existence results for the bistable and Fisher waves are based. The basic idea for the first theorem is quite simple:

1. Construct a homotopy of the $2n$-dimensional system to a system with an invariant two-dimensional subsystem on which the dynamics are those of the standard one species bistable problem.
2. Prove the existence of a wave speed for which the heteroclinic orbit occurs in the two-dimensional subsystem.
3. Construct a homotopy back to the original problem and conclude that for some wave speed the heteroclinic orbit, and hence the travelling wave exists for the original problem.

Theorems in this spirit are quite well known and are the basis for most applications of degree theory. In this latter setting a typical theorem goes as follows. Given a parameterized family of functions $f^\sigma : U \to \mathbf{R}$ such that the zeros of the functions do not lie on the boundary of U, then the degree of the function at one parameter value is the same as the degree at any other parameter value. Furthermore, if the degree is nonzero then for any σ there exists $x^\sigma \in U$ such that $f^\sigma(x^\sigma) = 0$. There are two key ingredients to this theorem:

(i) The zeros do not occur on the boundary, and

(ii) At some parameter value one can compute the degree.

Though, as was indicated in the introduction, our approach is based on the Conley index rather than degree theory, the spirit of the two methods have much in common.

Before stating the theorems let us establish some notation. Let $\phi : \mathbf{R} \times X \to X$ be a flow on a locally compact space. Given a compact set N, let

$$N^T = \{x \in N | \phi([-T,T],x) \subset N\}.$$

N is called an *isolating neighborhood* if there exists $T > 0$ such that $N^T \subset \text{int}(N)$. Let $N^\infty = \bigcap_{T>0} N^T$. It is easy to check that N^∞ is an invariant set, i.e., $\phi(\mathbf{R}, N^\infty) = N^\infty$. Clearly, N is an isolating neighborhood if and only if $x \in N^\infty$ implies $x \notin \partial N$. Notice that the role played by isolating neighborhoods in Conley index theory is equivalent to that played by the above mentioned set U in degree theory. An invariant set S is called *isolated* if there exists an isolating neighborhood N of S such that $S = N^\infty$.

A simple, but as we shall see, useful decomposition of an invariant set S is that of an *attractor-repeller* (A-R) pair. Let $\omega(U)$ and $\omega^*(U)$ denote the omega and alpha limit sets of U, respectively. $A \subset S$ is called an *attractor* in S if there exists a neighborhood U of A such that $\omega(U \cap S) = A$. The *dual repeller* of A, denoted by A^* is defined by $A^* = \{x \in S|\omega(x) \cap A = \emptyset\}$. The pair (A, A^*) make up an A-R pair. Notice that given an A-R pair decomposition of S, if $x \in S$, then $x \in A \cup A^*$ or $\omega(x) \subset A$ and $\omega^*(x) \subset A^*$, i.e., S is made up of the attractor A, its dual repeller A^*, and connecting orbits from A^* to A. Thus, if one lets $C(A^*, A)$ denote the set of connecting orbits from A^* to A, then $S = A \cup A^* \cup C(A^*, A)$.

Since the complicated $2n$-dimensional problem will be mapped by a homotopy to a simple two-dimensional problem we need to discuss what is meant by continuing isolating neighborhoods and isolated invariant sets. Consider a parameterized family of flows $\phi(t,x,\sigma) = \phi^\sigma(t,x)$, where $\sigma \in [0,1]$ is the parameter value. Define the parameter flow to be $\Phi : \mathbf{R} \times X \times [0,1] \to X \times [0,1]$ by

$$\Phi(t,x,\sigma) = (\phi^\sigma(t,x),\sigma).$$

Let N^σ denote an isolating neighborhood for the flow ϕ^σ. One says that the isolating neighborhood N^1 continues to N^0 if there exists N an isolating neighborhood of Φ, the parameter flow, such that $N|_{X\times\{0\}} = N^0$ and $N|_{X\times\{1\}} = N^1$. Similarly, if S^σ denotes an isolated invariant set under ϕ^σ, then S^0 continues to S^1 if there exist corresponding isolating neighborhoods that continue.

We are now ready to present the hypotheses for the abstract theorem, and start with the homotopy discussed in the previous paragraph. With this in mind we write down the following two parameter set of equations:

$$(\mathrm{TW}^{c,\sigma}) \qquad \begin{aligned} u' &= p, \\ M(\sigma)p' &= -cp - H(u,\sigma), \end{aligned}$$

where $u,p \in \mathbf{R}^n$, $M(\sigma)$ is a diagonal matrix with positive entries μ_i^σ, $c \in \mathbf{R}$ is the wave speed, $\sigma \in [0,1]$ is the homotopy parameter, and $H : \mathbf{R}^n \times [0,1] \to \mathbf{R}^n$. The corresponding family of reaction systems is given by

$$(\mathrm{R}^\sigma) \qquad u' = H(u,\sigma).$$

We shall adopt the convention that at $\sigma = 0$ we are at the original system we wish to study, i.e., $H(u,0) = h(u)$ and $\mu_i^0 = \mu_i$, and at $\sigma = 1$ we have the simple two-dimensional invariant subsystem for which we can prove the existence of a travelling wave solution. We formalize this via the following assumption.

Assumption I. For the reaction system (R^1) there exists a one-dimensional attracting invariant subspace $L \subset \mathbf{R}^n$. Furthermore, the dynamics on L are as in Fig. 2, i.e., the set of bounded solutions consist of the hyperbolic critical points $\{A,B,C\}$ and the heteroclinic connections $C \to A$ and $C \to B$.

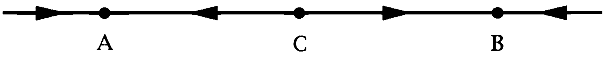


FIG. 2. *The dynamics on L.*

We now state the assumption concerning the homotopy from $\sigma = 1$ to $\sigma = 0$. Let $A_0^1 = (A^1,0), B_0^1 = (B^1,0)$, and $C_0^1 = (C^1,0)$.

Assumption II. A^σ and B^σ continue as critical points for (R^σ) and N^σ continue as isolating neighborhoods for $(\mathrm{TW}^{c,\sigma})$ such that (B_0^σ, A_0^σ) is an attractor-repeller pair for $(N^{c,\sigma})^\infty$, the invariant set isolated by N^σ under the flow generated by $(\mathrm{TW}^{c,\sigma})$. Furthermore, for $\sigma = 1$, if there exists an $A_0^1 \to B_0^1$ solution to $(\mathrm{TW}^{c,1})$ for some value of c, then the connecting orbit belongs to $(N^{c,1})^\infty$.

THEOREM 3.1. *Given Assumptions* I *and* II, *there exists a bistable travelling wave from* A_0 *to* B_0 *for some wave speed* $\bar{c}$.

The reader is referred to the appendix for the technical aspects of the proof of this theorem. The general idea, however, is quite simple. Assumption I is used to guarantee that the Conley index (as a connected simple system) differs for $c \approx +\infty$ and $c \approx -\infty$. Now Assumption II, along with an a priori bound guarantees that the index remains the same throughout the homotopy for $|c|$ sufficiently large. Of course this implies that the index at $c^0 \approx +\infty$ and $c^0 \approx -\infty$ differ, and hence there exists an A_0^0 to B_0^0 connection for some intermediate wave speed $\bar{c} = \bar{c}^0$.

With regard to the existence of a Fisher wave we have the following result.

THEOREM 3.2. *Let A and C be hyperbolic critical points for the reaction system* (2.2)*, where A is stable and C has a one-dimensional unstable manifold. If there exists a $C \rightarrow A$ heteroclinic orbit, a solution to* (2.2)*, then there exists a $\hat{c}$ such that for all $c \in (\hat{c}, \infty)$ there is an $A_0 \to C_0$ connecting orbit for the travelling wave system* (2.3).

Again the proof of this theorem appears in the appendix; however, we remark at this point that it follows directly from results of Conley and Fife [3] and McCord [16].

4. Existence of a bistable wave. We shall show that a straightforward application of the abstract Theorem 3.1 yields the existence of a monotone A_0 to B_0 connection for the system (2.3), and so, a bistable travelling wave for the reaction-diffusion system (1.1). In particular, we shall prove the following theorem.

THEOREM 4.1. *Under hypotheses* (H1)–(H5)*, there exists a wave speed $\bar{c}$ for which there is a monotone $A_0 \to B_0$ connection for* (2.3).

Of course, to apply Theorem 3.1, one needs to demonstrate that Assumptions I and II are satisfied. This shall be done in several steps beginning with the construction of the homotopy F. Let

$$\phi_i = \{u | f_i(u) = 0\},$$

so that ϕ_i represents the zero surface of f_i. By the assumption (H4)

$$\bigcap_{i=1}^{n} \phi_i = \{B, C\}.$$

Let L denote the diagonal of $\mathbf{R}^n$. Since $B \in L$, there exists an index j such that ϕ^j intersects L transversally and such that this intersection contains at least two points. By a reordering of the indices, we arrange that $j = 1$, and from (H4) we may assume without loss of generality that this intersection has exactly two points. For, if there are more points in the intersection, then we may homotope ϕ^1 to eliminate the extra points of intersection; in view of (H4) this can be done without changing the number of critical points. Let $\rho = f^1$, and define

$$\begin{aligned} F_i(u, \sigma) &= (1-\sigma) f_i + \sigma \rho(u_i, u_2, \ldots, u_{i-1}, u_1, u_{i+1}, \ldots, u_n), \\ \mu_i^\sigma &= \mu_i(1-\sigma) + \sigma\,. \end{aligned}$$

For $\sigma \in [0, 1]$ consider the system

$$\text{(TW}(c,\sigma)) \qquad \begin{aligned} &(4.1\text{a}) & u_i' &= p_i, \\ &(4.1\text{b}) & \mu_i^\sigma p_i' &= -cp_i - u_i F_i(u, \sigma). \end{aligned}$$

Obviously, this is a particular choice for the general system ($\text{TW}^{c,\sigma}$). As σ varies, the position of the intermediate critical point varies; let this be denoted by C^σ for the reaction system and C_0^σ for the travelling wave system.

The second step in applying Theorem 3.1 is to construct an isolating neighborhood. The fact that we are looking for monotone $A \rightarrow B$ connecting orbits strongly suggests how the isolating neighborhood should be chosen. In particular, monotonicity implies two things; $0 \le u_i(t) \le 1$, and $0 \le p_i(t)$ for $t \in \mathbf{R}$. This suggests choosing a rectangular box, $\mathcal{R} = \mathcal{R}_u \times \mathcal{R}_p$, where $\mathcal{R}_u = [0,1]^n = \bar{R}$ and $\mathcal{R}_p = [0,K]^n$ for some sufficiently large K. Clearly, if $A \to B$ is monotone then $A_0 \to B_0 \subset \mathcal{R}$. On the other hand, it is equally obvious that $\mathcal{R}$ is not an isolating neighborhood since the critical points are elements of $\partial\mathcal{R}$. This suggests defining

$$N^\sigma = \mathcal{R} \cup \bar{B}_\epsilon(A_0) \cup \bar{B}_\epsilon(B_0) \backslash B_\epsilon(C_0^\sigma),$$

where $B_\epsilon(P)$ and $\bar{B}_\epsilon(P)$ denote the open and closed balls, respectively, with center P and radius ϵ. With this choice of N^σ we can now verify that Assumption I is satisfied.

LEMMA 4.2. *Assumption* I *is satisfied by the equations* (4.1).

Proof. It is clear from the symmetry in the construction of $F(\cdot, 1)$ that L is an invariant subspace of $\mathbf{R}^n$ under the reaction system R^1, and that the zero sets of $F_i(\cdot, 1)$ intersect L transversally at exactly two points, B and C^σ. It thus follows from (H4) and (H5) that Assumption I holds. □

The main burden of the proof is to show that Assumption II is satisfied, and while not overwhelmingly difficult, does involve several rather technical steps. Recall from the discussion at the beginning of §3, that N being an isolating neighborhood is equivalent to showing that $N^\infty \cap \partial N = \emptyset$. In Lemma 4.3 we will show that (B_0, A_0) form an attractor-repeller pair for $(N^{c,\sigma})^\infty$, the invariant set in N^σ under the flow induced by $(\mathrm{TW}^{c,\sigma})$. Thus, $(N^{c,\sigma})^\infty = A_0 \cup B_0 \cup C(A_0, B_0)$. Since $A_0 \cup B_0 \subset \mathrm{int}(N^\sigma)$, we only need to demonstrate that $(A_0 \to B_0) \cap \partial N^\sigma = \emptyset$. This will be done via a series of propositions and lemmas. To see what is involved notice that ∂N^σ is a subset of the points in

$$(\partial\mathcal{R}_u \times \mathcal{R}_p) \cup (\mathcal{R}_u \times \partial\mathcal{R}_p),$$
$$\partial(B_\epsilon(A_0) \cup B_\epsilon(B_0)) \backslash \mathcal{R},$$

and

$$\partial(B_\epsilon(C_0^\sigma)) \cap \mathcal{R}.$$

Thinking of N^σ as a "box" and of the p_i's as defining "vertical directions," we refer to $\partial\mathcal{R}_u \times \mathcal{R}_p$ as the sides of N^σ, elements of $\mathcal{R}$ where $p_i = 0$ for some i as lying on the bottom of N^σ, and elements of $\mathcal{R}$ where $p_i = K$ for some i as lying on the top of N^σ. Lemmas 4.4 and 4.5 assert that an $A_0 \to B_0$ orbit cannot touch the top and bottom of the box respectively. Next we prove Proposition 4.6 which guarantees that there is no tangency to the small ball around the critical point C_0. Finally, Proposition 4.7 implies that if $u \in A_0 \to B_0$, then $u \in \mathcal{R}_u$, i.e., $A_0 \to B_0$ does not intersect the sides of $\mathcal{R}$ nor $\partial(B_\epsilon(C_0^\sigma)) \cap \mathcal{R}$.

LEMMA 4.3. (B_0, A_0) *forms an attractor repeller pair for* $(N^{c,\sigma})^\infty$ *under the flow generated by* $(\mathrm{TW}^{c,\sigma})$.

Proof. Since $p_1 \geq 0$ for all $(u,p) \in N^\sigma \backslash (B_\epsilon(A_0) \cup B_\epsilon(B_0))$, $-u_1$ acts as a Lyapunov function for the flow on this subset of N^σ. On the other hand, if ϵ is sufficiently small, then the fact that A_0 and B_0 are hyperbolic critical points insures, via the Hartman–Grobman theorem, that the only bounded orbits in $B_\epsilon(A_0) \cup B_\epsilon(B_0)$ are the critical points themselves. □

LEMMA 4.4. *There exists* K *such that on any orbit in* $(N^{c,\sigma})^\infty$, $p_i < K/2$ *for* $(\sigma, c) \in [0,1] \times \mathbf{R}$ *and all* i.

Proof. Since any such orbit has $p_i = 0$ at its endpoints, at a maximum of p_i, $dp_i/du_i = 0$. Therefore, from (4.1b), $p_i = -c^{-1}u_i F_i(u, \sigma)$. This provides a bound on p_i for $|c| \geq 1$.

Consider next $|c| < 1$, and with $H_i(u, \sigma) = u_i F_i(u, \sigma)$, set

$$m = \max_i \max_{\bar{R}} |H_i(u, \sigma)|.$$

Suppose that $p_i \geq 0$ for $u_i \in [u_{i0}, u_{i1}]$ and that it vanishes at these points. We claim that

$$p_i < 2\mu_i^{-1}[|c| + (c^2 + 2\mu_i m)^{1/2}] = p_{M_i}$$

say. For, if this is false for some i and u_i, then $p_i(u_i) = p_{M_i}$. Then from (2.5) and the mean value theorem for integrals, since $p_i(u_{i0}) = 0$,

$$\tfrac{1}{2}\mu_i p_{M_i}^2 \leq |c| p_{M_i} + m.$$

Hence

$$p_{M_i} < \mu_i^{-1}[|c| + (c^2 + 2\mu_i m)^{1/2}],$$

a contradiction. It is thus clear that a choice of K with the stated properties is possible.

LEMMA 4.5. *Suppose $(u,p) \notin \bar{B}_\epsilon(A_0) \cup \bar{B}_\epsilon(B_0)$. If $p_i = 0$ for some i, then $(u,p) \notin (N^{c,\sigma})^\infty$.*

Proof. Under the stated conditions, if $(u,p) \in (N^{c,\sigma})^\infty$ then $p_i = p_i' = 0$ and $p_i'' \geq 0$. Differentiating (4.1b) we obtain

$$\mu_i^\sigma p_i'' = -\sum_{j \neq i} \frac{\partial F_i}{\partial u_j} u_j' < 0,$$

unless $p_j = u_j' = 0$ for all $j = 1, \ldots, n$. In this case $p_j = p_j' = 0$ for all $j = 1, \ldots, n$, so from (4.1b) the point is a critical point. □

PROPOSITION 4.6. *Orbits in $(N^{c,\sigma})^\infty$ are not tangent to $\bar{B}_\epsilon(C_0) \cap \partial N^\sigma$ for small enough ϵ.*

Proof. The proof follows from examining the linearized operations about the critical point C_0^σ. What needs to be shown is that a monotone orbit from A_0 to B_0 cannot be tangent to a small enough ball centered at C_0^σ. Observe that if all the eigenvalues are complex then the solutions near C_0^σ will oscillate, and hence, exit N^σ. Thus, if there is a tangency, both of the following must hold: there must be a real nonnegative eigenvalue, and a real nonpositive eigenvalue with corresponding eigenvectors, respectively, each of the form (u,p) with $u_i > 0$ for all $i = 1, \ldots, n$ (if not, then the eigenvector points out of N^σ). We shall rule out these possibilities by a contradiction argument. So suppose there exist $c \geq 0$ and an eigenvalue $\lambda \geq 0$ with associated eigenvector having $u_i > 0$ for all i. Linearization leads to the system

$$-\lambda u_i + p_i = 0, \tag{4.2}$$

$$-\Sigma a_{i_k} u_k - (c + \lambda\mu_i)p_i = 0. \tag{4.3}$$

Substitution of the first equation in the second yields the relation

$$T_i u_i + \sum_{k \neq i} a_{ik} u_k = 0, \tag{4.4}$$

where $T_i = (c + \lambda\mu_i)\lambda + a_{ii}$. Of course from (H2), $a_{ii} < 0$, $a_{ij} > 0\,(i \neq j)$ for any i. It follows that $T_i < 0$ for each i, as the summation term in (4.4) is strictly positive. Thus $a_{ii} \leq T_i < 0$. Note finally that for some i, $\Sigma a_{i_k} u_k > 0$, as otherwise (H5) is contradicted, the matrix $[a_{ij}]$ being nonsingular by (H4). Hence

$$-a_{ii}u_i < \sum_{k \neq i} a_{i_k} u_k = -T_i u_i\,.$$

This contradicts the above inequality, and shows that if $c \geq 0$, an eigenvector with $u_i > 0$ (for all i) cannot be associated with a nonnegative eigenvalue. When $c \leq 0$, an analogous argument holds for a nonpositive eigenvalue. □

PROPOSITION 4.7. *Let $(u,p) \in (N^{c,\sigma})^\infty \backslash (A_0 \cup B_0)$, then $u_i \in (0,1)$.*

The proof of this proposition entails checking the travelling wave near the critical points A_0 and B_0 and along the sides of R. The following lemma will prove useful.

LEMMA 4.8. *Let (u,p) be a solution to $(\mathrm{TW}^{c,\sigma})$ and assume that $u_i(t_0)$ is a local maximum (minimum) of $u_i(t)$, then $f_i(u,\sigma) > 0$ (< 0).*

Proof. If $u_i(t_0)$ is a local maximum (minimum) of $u_i(t)$, then $u_i' = p_i = 0$. Now consider the second derivative

$$\begin{aligned} p_i' &= \mu_i^{-1}(-cp_i - u_i f_i(u,\sigma)) \\ &= \mu_i^{-1}(-u_i f_i(u,\sigma)) \end{aligned}$$

Thus $u_i'' < 0$ if and only if $f_i(u,\sigma) > 0$. □

We shall now show that away from the critical points $A_0 \rightarrow B_0$ does not intersect the sides of R.

LEMMA 4.9. *If $(u,p) \in (N^{c,\sigma})^\infty \backslash (B_\epsilon(A_0) \cup B_\epsilon(B_0))$, then $u_i \in (0,1)$.*

Proof. There are two cases to consider; $u_j = 0$ and $u_j = 1$ for some $j = 1, \ldots, n$. So assume that $u_j = 0$, then $(u,p) \in \partial N^\sigma \backslash (B_\epsilon(A_0) \cup B_\epsilon(B_0))$. $(u,p) \in (N^{c,\sigma})^\infty$ implies that $p_j = 0$ (otherwise (u,p) leaves N^σ in forward or backward time). But, as is easy to check, $\{(u,p) | u_j = p_j = 0\}$ is an invariant subspace. Thus (u,p) must lie on a bounded orbit on this hyperplane contradicting the fact that $(u,p) \in A_0 \rightarrow B_0$.

Thus, we can assume that $u_j = 1$ and $p_j = 0$. Note that $F_j(u,\sigma) < 0$, if $u \in \mathcal{R} \backslash B_0$ and $u_j = 1$. By Lemma 4.8, this implies that u_j attains a local minimum at this point, contradicting the monotonicity assumption. □

Notice that we have reduced the problem to studying the trajectory $A_0 \rightarrow B_0$ in the regions $B_\epsilon(A_0)$ and $B_\epsilon(B_0)$. The following lemma shows that as $t \rightarrow -\infty$, $u_i(t)$ remains positive for $i = 1, \ldots, n$.

LEMMA 4.10. *If $(u,p) \in (N^{c,\sigma})^\infty \cap B_\epsilon(A_0)$, then $u_i, p_i > 0$, for $i =, \ldots, n$.*

Proof. We first show that $p_i > 0$ for each $i = 1, \ldots, n$. Note first that from the proof of Lemma 4.5 there cannot be a point on the orbit where $p_i = p_i' = 0$. From (H1)–(H5) there is a choice of $\epsilon > 0$ such that the p_i equation may be written in the form

$$\mu_i p_i' = -cp_i + k_i u_i[1 + \delta(u)], \tag{4.5}$$

with $k_i > 0$ and $|\delta(u)| < \frac{1}{2}$ for $|u| \leq \epsilon$. Assume that an orbit in $(N^{c,\sigma})^\infty$ cuts the boundary of $B_\epsilon(A_0) \cap \mathbf{R}_+^n$ at $t = 0$. From the definition, it must necessarily converge to A_0 as $t \rightarrow -\infty$. Now suppose that as t decreases from zero on this orbit $p_i = 0$ first at u_i. Then it follows from (4.5) that if $u_i > 0$ ($u_i < 0$, respectively), u_i has a strict minimum (respectively, a strict maximum). In the first case, the orbit clearly cannot approach A_0 without another turning point. But such a turning point can only be a minimum. So this possibility is ruled out. In the second case, a similar argument shows that in fact, the orbit cannot cut the boundary of $B_\epsilon(A_0) \cap \partial\mathbf{R}_+^n$. Hence, p_i cannot vanish. Obviously, then $u_i > 0$ also, for otherwise it must have a minimum at which $p_i = 0$, which was ruled out above. □

The proof of Proposition 4.7 will be complete once we show the following.

LEMMA 4.11. *If $(u,p) \in (N^{c,\sigma})^\infty \cap B_\epsilon(B_0)$, then $u_i < 1$, for $i, \ldots, n$.*

The proof of Lemma 4.10 was made easy by the fact that the hyperplanes $u_i = 0$ define invariant subspaces for the travelling wave system. Of course, near the critical point B_0 we do not have such nice control on the invariant subspaces, thus the proof of Lemma 4.11 is more complicated. The assumption (H4) is important at this point. Since B_0 is a hyperbolic fixed point Hartman–Grobmann implies that we can reduce the problem to a careful study of the flow of the linearized system about B_0. To keep the notation to a minimum we first translate the origin to $(1,\dots,1)$. With this in mind we make the following definitions. For fixed σ let $H : \mathbf{R}^n \to \mathbf{R}^n$ be defined by $H(u) = (u_1F^1(u,\sigma),\dots,u_nF^n(u,\sigma))$. The linearized version of $(\mathrm{TW}^{c,\sigma})$ at the point B_0 is given by

$$\begin{aligned} u' &= p, \\ Mp' &= -cp - DH(B_0)u. \end{aligned} \tag{4.6}$$

To further simplify the notation we shall write this system as

$$\begin{aligned} u_i' &= p_i, \\ \mu_i^\sigma p_i' &= -cp_i - g_i(u), \end{aligned} \tag{4.7}$$

where $g_i(u) = \nabla F_i(B,\sigma)\cdot u$. Define $\mathbf{e}^i \in \mathbf{R}^n$ by

$$\mathbf{e}^i \cdot \nabla F_j(B,\sigma) = \delta^i_j.$$

Define $\rho = (\rho_1,\dots,\rho_n)$, where $\rho_i \in \{0,\pm 1\}$. Now define the closed cones

$$\begin{aligned} K(\rho) &= cl\left\{u = \sum \alpha_i e^i | \alpha_i = \rho_i = 0 \text{ or } \alpha_i\rho_i > 0\right\}, \\ Q(\rho) &= cl\{u|u_i = \rho_i = 0 \,\text{or}\, u_i\rho_i > 0\}. \end{aligned}$$

By Remark 2.1, $K(-1,-1,\dots,-1) \subset Q(1,1,\dots,1)$. Consider Fig. 3 which describes the zero sets for g_i and the cones $K(\rho)$ and $Q(\rho)$ for $n = 2$. Using it and the linearized system we shall now give a simple proof of Lemma 4.11 for the special case $n = 2$.

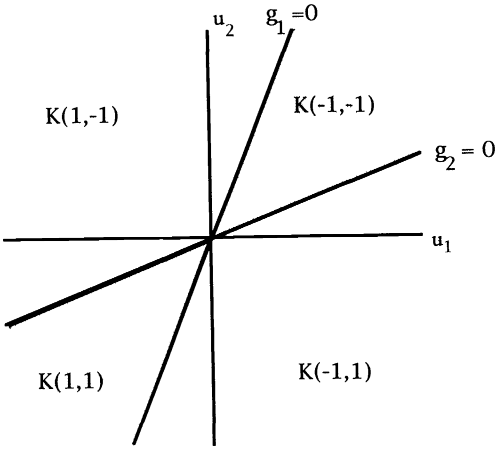


FIG. 3

LEMMA 4.12. *Let $n = 2$, and let $(u(t),p(t))$ be a solution to (4.7) such that $u(0) \in Q(-1,-1)$, $p(0) \in Q(1,1)$, and $\lim_{t\to\infty}(u(t),p(t)) = (0,0)$. Then $u([0,\infty)) \subset Q(-1,-1)$.*

Proof. Assume not, then there exists a first time t_0 such that $u(t_0) \in \partial Q(-1,\ldots,-1)$ and $p(t_0) \in Q(1,\ldots,1)$. Without loss of generality we can, therefore, assume that $u_1(t_0) = 0$ and $p_1(t_0) > 0$. Since $\lim_{t\to\infty} u_1(t) = 0$, there exists $t_1 > t_0$ such that $u_1(t_1)$ is a local max of u_1. Now by Lemma 4.8 and the fact that $u_1(t_1) > 0$ we have that

$$u(t_1) \in K(1,*) \cap Q(1,*),$$

where $*$ should be replaced by either $+1$ or -1. From Fig. 3, it is clear that this implies that

$$u(t_1) \in K(1,-1) \cap Q(1,1).$$

In turn this implies that there exists $t_2 > t_0$ such that $u_2(t_2)$ is a local maximum of u_2. Since $K(1,1) \subset Q(-1,-1)$, it is clear that $t_1 \neq t_2$. So assume that $t_2 > t_1$. Then $u(t_2) \in K(-1,1)$ and for all $t > t_2$, $u(t) \in K(-1,1)$. This contradicts the existence of t_1. Assuming that $t_1 > t_2$ leads to a similar contradiction. □

Notice that in the proof of Lemma 4.12 it was essential that $K(-1,-1) \subset Q(1,1)$, and thus, separated the cones $K(1,-1)$ and $K(-1,1)$ in $Q(1,1)$. The proof of Lemma 4.11 will now follow by induction on n the number of species. But first we prove a simple lemma.

LEMMA 4.13. *Let $\hat{u}_n > 0$ and define $\hat{u} = (\hat{u}_1,\ldots,\hat{u}_n)$ by $g_i(\hat{u}) = 0$ for $i = 1,\ldots,n-1$. Then $\hat{u}_i > 0$ for all i.*

Proof. Notice that $\hat{u} = \theta \mathbf{e}^n$, for some $\theta \in \mathbf{R}$. By Remark 2.1, $e_i^n < 0$ for $i = 1,\ldots,n$. Now $\hat{u}_n > 0$ implies $\theta < 0$, and hence, $\hat{u}_i > 0$ for $i = 1,\ldots,n$. □

Proof of Lemma 4.11. The proof is by induction on n the dimension of the reaction system. Lemma 4.12 guarantees that the result is true for $n = 2$. So assume that the result is true for $n \leq \bar{n}$. We need to show that it holds for $n = \bar{n} + 1$.

Let $\gamma = \cup_{t\geq 0} u(t)$. There are two cases to consider. First, there exists i, such that $\gamma \subset \cup_{\rho_i=-1} Q(\rho)$, and second, such an i does not exist. If we are in the first case, then without loss of generality we can assume that $i = n$. Now define $\bar{\gamma} = \cup_{t\geq 0}(u_1(t),\ldots,u_{n-1}(t),0)$, i.e., the projection of γ onto the subspace $\mathbf{R}^{n-1} \times \{0\}$. By the induction hypothesis $\bar{\gamma} \subset Q(-1,\ldots,-1) \subset \mathbf{R}^{n-1}$ and the lemma holds.

Now consider the second possibility. The argument which follows is probably clearer if one keeps Fig. 4 in mind (this is the figure corresponding to $n = 3$). In this case there exists t_1 such that $u_n(t_1) > 0$, and hence there exists

$$\hat{u}_n = \max_{t>0} u_n(t) > 0.$$

Let us choose t_1 such that $u_n(t_1) = \hat{u}_n$. Now define $\hat{\gamma}$ to be the projection of γ onto the space $\mathbf{R}^{n-1} \times \{\hat{u}_n\}$. For $i = 1,\ldots,n-1$, define $\hat{u}_i$ by $g_i(\hat{u}_i) = 0$ and let $\hat{u} = (\hat{u}_1,\ldots,\hat{u}_n)$. Since we want to work on the subspace $\mathbf{R}^{n-1} \times \{\hat{u}_n\}$ define $\hat{\mathbf{e}}^i = (e_1^i,\ldots,e_{n-1}^i,0)$. Similarly, define $\hat{K}(\rho)$ as before but in terms of the $\hat{\mathbf{e}}^i$'s and $\hat{Q}(\rho)$ as the orthants in $\mathbf{R}^{n-1}\times\{\hat{u}_n\}$ based at the point $\hat{u}$. By the induction hypothesis, it is clear that $\hat{\gamma} \subset \hat{Q}(-1,\ldots,-1)$. Of course, this does not yet prove the lemma, since by Lemma 4.13 $\hat{u}_i > 0$, and thus, $\hat{Q}(-1,\ldots,-1) \not\subset Q(-1,\ldots,-1)$. However, $u(t_1) = \hat{u}+w$ where $w_i \leq 0$ for $i = 1,\ldots,n-1$, and w_n=0. Now, $\hat{u} = \theta\mathbf{e}^n$ where $\theta < 0$; hence $g_n(\hat{u}) < 0$. Furthermore, $w \cdot \mathbf{e}^n < 0$, thus $g_n(u(t_1)) = g_n(\hat{u} + w) < g_n(\hat{u}) < 0$. Therefore, $u(t_1) \notin K(*,\ldots,*,1)$. Of course, this contradicts the assumption that at t_1, u_n has a local maximum. □

5. Existence of a Fisher wave. As an application of Theorem 3.2, we show that for large c there is an $A_0 \to C_0$ connection. Unfortunately, the argument does not

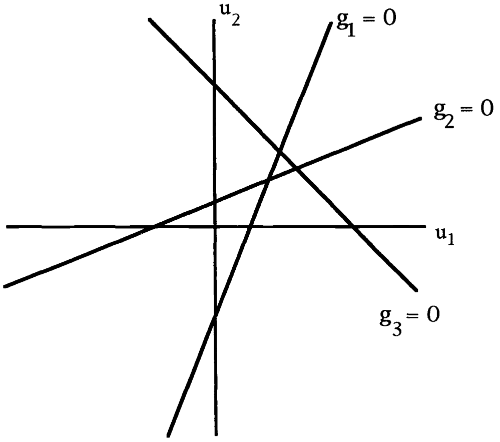


FIG. 4. *Zero sets for g_i on the plane $\mathbf{R}^2 \times \{\hat{u}_n\}$ where $\hat{u}_n > 0$.*

show that this is a monotone connection. However, for two species we can establish the occurence of a monotone $A_0 \to C_0$ connection, i.e., a Fisher wave. The difficulty in extending this result to n species is that we are not able to deal with the local analysis near C_0 to show that there is no oscillation in $B_\epsilon(C_0)$.

THEOREM 5.1. *For large enough c, there exists an $A_0 \to C_0$ connection, and if $n = 2$, there exists a monotone $A_0 \to C_0$ connection (that is, there is a Fisher wave).*

Proof. The first part of this theorem follows directly from (H4), Remark 2.3, and the application of Theorem 3.2.

The stronger result for the case $n = 2$ can be obtained as follows. In close analogy with the proof of Theorem 4.1, define

$$N = \{(u,p) \mid 0 \le u_i \le c_i; 0 \le p_i \le K\} \cup \bar{B}_\epsilon(A) \cup \bar{B}_\epsilon(C).$$

With the exception of the dynamics in $B_\epsilon(C_0)$, the relevant lemmas and propositions of the previous section are still applicable. Thus, to show that N is an isolating neighborhood with attractor-repeller pair (C_0, A_0), it is clearly enough to rule out tangencies to $B_\epsilon(C_0)$. Furthermore, by Lemma 4.10 we have monotonicity in $B_\epsilon(A_0)$, and hence, showing monotonicity in $B_\epsilon(C_0)$ will establish the result. As C_0 is hyperbolic, the flow in $B_\epsilon(C_0)$ may be approximated by the flow generated by the linearization. We shall show that all the associated eigenvalues are real, and that the eigenvector corresponding to the negative eigenvalue of smallest magnitude is of the form $(u_1, u_2, p_1, p_2) = (+, +, -, -)$. This is enough to prove the result.

With $A_{ij} = c_i \partial f_i / \partial u_j$, from assumption (H2) (see also Fig. 1), $a_{ii} < 0, a_{ij} > 0$ for $(i \ne j)$, and $\Delta = a_{11}a_{22} - a_{12}a_{21} < 0$. Linearization leads to the matrix

$$\begin{vmatrix} -\lambda & 0 & 1 & 0 \\ 0 & -\lambda & 0 & 1 \\ -\frac{a_{11}}{\mu_1} & -\frac{a_{12}}{\mu_1} & -\frac{c+\lambda\mu_1}{\mu_1} & 0 \\ -\frac{a_{21}}{\mu_2} & -\frac{a_{22}}{\mu_2} & 0 & -\frac{c+\lambda\mu_2}{\mu_2} \end{vmatrix},$$

and the eigenvalues λ are the solutions of

$$F(\lambda) \overset{\text{def}}{=} [\lambda(\lambda\mu_1 + c) + a_{11}]\,[\lambda(\lambda\mu_2 + c) + a_{22}] - a_{12}a_{21} = 0.$$

We claim that for c large enough positive, there are three negative real roots $\lambda_1, \lambda_2, \lambda_3$ with $\lambda_3 \leq \lambda_2 < \lambda_1$, and one positive root. Also, $-c/2\mu < \lambda_1 < 0$ where $\mu = \max(\mu_1, \mu_2)$. In the following we assume without loss of generality that $\mu_1 > \mu_2$.

The existence of one positive root is obvious, for $F(0) = \Delta < 0$. We have

$$F(-c/\mu_2) = \Delta - c^2 a_{22}(\mu_2 - \mu_1)/\mu_2^2 < 0.$$

Also,

$$16\mu_1^3 F\left(-\frac{c}{2\mu_1}\right) = [c^2 - 4\mu_1 a_{11}]\,[c^2(2\mu_1 - \mu_2) - 4\mu_1^2 a_{22}] - 16\mu_1^3 a_{12}a_{21} > 0$$

for large enough c. These inequalities establish the claim.

Finally, to find the direction of the eigenvector ϕ associated with λ_1, we note that from the matrix it is of the form $\phi = (u_1, u_2, \lambda_1 u_1, \lambda_1 u_2)$. Substitution in the third equation then yields

$$u_1\left[-\frac{c^2}{2\mu_1} + a_{11}\right] + a_{12}u_2 = 0,$$

and taking c large, we obtain the result. □

6. Travelling waves of invasion and dominance. From the point of view of applications it is important to discover whether a bistable wave with negative speed can exist; such a wave may be described as a "wave of invasion" since the species migrate into a spatial region where their initial density is low. Since the mutualism is obligate, so that all species must be present for survival, it is not immediately clear whether such a wave can exist. From another point of view, this question can be phrased in terms of the relative dominance of the equilibria A and B; see [6] for a discussion in the one species case, and [4]. It is natural to define B as dominant (over A) if a monotone increasing travelling wave has negative wave speed, for then the solution u of the reaction-diffusion system (1.1) will approach B on compact sets. The following result gives a condition for dominance of B and also says something about the basin of attraction; a complementary result for the dominance of A may be obtained in a very similar manner. The result is weaker than the necessary and sufficient condition that may be obtained for the one species case, but it seems likely that such a condition may be very hard to find in the present case.

We start by defining a function ϕ which is used to construct a subsolution. Let $M \in C([0,1], \mathbf{R})$ with $M(0) = 0$, and consider the equation

$$\phi'' + M(\phi) = 0 \tag{6.1}$$

with $\phi'(0) = 0$. Put $V(u) = \int_0^u M(s)ds$, and assume that $V(1) > 0$. A simple energy argument shows that for some $b \in [0,1]$, the solution of (6.1) with $\phi(0) = b$ reaches zero (obviously symmetrically) at some time t_0. Define $\phi(t) = 0$ for $|t| > t_0$.

LEMMA 6.1. *Define*

$$M(u) = \min_{1 \leq i \leq n} \mu_i^{-1} h_i(u_1, \dots, u_n),$$

and suppose that $V(1) > 0$. Then $\underline{\phi} = (\phi, \dots, \phi)$ is a subsolution for the reaction-diffusion system (1.1), and a solution $u(x,t)$ with $u(x,0) \geq \underline{\phi}(x)$ tends to B uniformly on compact sets. Furthermore, a bistable wave exists, and its speed $c < 0$.

Proof. The assertions concerning the subsolution follow directly from [13]. From Theorem 4.1 a bistable wave exists. Necessarily, its speed c is less than zero, for otherwise a contradiction would be obtained by translating the travelling wave so that at $t = 0$ it is above the subsolution. □

The speed of travelling waves is also of practical importance. For two species, if $\mu_1 \ll \mu_2$, there are formal arguments which suggest that invading waves are "slow" in the sense that $|c|$ is proportional to $\mu_1^{1/2}$, whereas "dying out" waves are fast with c proportional to $\mu_2^{1/2}$. However, the following result is the only rigorous result known to us.

LEMMA 6.2. *Let ρ_1 be positive with*

$$\rho_1^2 = 4 \min_{1\le i\le n} \max_{u\in \bar{R}} h_i(u). \tag{6.2}$$

For a bistable wave with $c < 0$, we have $|c| \le \rho_1$.

Proof. For convenience put $\rho = -c$. We shall show that if $\rho > \rho_1$, for a monotone orbit in $\bar{R}$, $p_i > 0$. Assume contrary to the assertion that p_i is the first of the p's to vanish and p_i has its first zero at $u_i = \alpha \in [0, b_i]$. On $u_i \in [0, \alpha]$ we may parameterize the orbit with u_i, and write $H(u_i) = h_i(u)$. Let L be the line $p_i = \rho u_i/2\mu_i$ in the $u_i - p_i$ plane. Near $u = 0$, $H(u) < 0$, so from (2.3), $\mu_i(dp_i/du_i) > \rho$, and $p_i > \rho u_i/\mu_i$. Thus the orbit lies initially above L. Furthermore, since $p_i(\alpha) = 0$, the orbit must cut L, first, say, at $(\beta, \rho\beta/2\mu_1)$. So from (2.3) again,

$$\begin{aligned}\mu_i \frac{dp_i}{du_i} &= p - \frac{2\mu_i H(\beta)}{\rho\beta}\\ &\ge \rho - \frac{2\mu_i}{\rho} \max\left[\frac{H(\beta)}{\beta}\right]\\ &> \frac{\rho}{2}\end{aligned}$$

from (6.2). However, as the orbit cuts L at β, $dp_i/du_i < \rho/2\mu_i$ there. This is a contradiction. □

7. Stability and uniqueness[2]**.** We consider here the stability of a bistable wave. Partial results on uniqueness are also obtained. The main tool in the investigation is a comparison principle for pseudomonotone systems analogous to that commonly used for a scalar diffusion equation; its validity here rests on the mutualistic assumption (H2).

Let U be an open neighborhood of $\bar{R}$. A smooth function $\underline{u} : \mathbf{R} \times [0, \infty) \to U$ is said to be a *subsolution* of (1.1) if

$$\frac{\partial \underline{u}_i}{\partial t} \le \mu_i \frac{\partial^2 \underline{u}_i}{\partial x^2} + h_i(\underline{u}).$$

A *supersolution* $\bar{u}$ is defined by reversing the inequality.

THEOREM 7.1 [24, §32]. *Let $\underline{u}$ and $\bar{u}$ be a subsolution and a supersolution, respectively, of* (1.1) *with $\underline{u}(x, 0) \le \bar{u}(x, 0)$. Then $\underline{u}(x, t) \le \bar{u}(x, t)$ for $x, t \in \mathbf{R} \times (0, \infty)$.*

Note that it is a consequence of our assumptions (H1), (H2), (H4), and (H5) that there is an $\eta \in \mathbf{R}^n$ with strictly positive components, such that the derivative in

[2] This type of result; generalizing the single species result in [7], seems first to have been considered in [15] and [10]. The version here is essentially due to Paul Fife.

the direction η of each f_i at B is negative. This will be used in the discussion of a subsolution which follows.

Let $u^*(x-ct)$ be a bistable travelling wave, and in (1.1) change to the travelling wave coordinate $z = x - ct$ obtaining the system

$$\frac{\partial u_i}{\partial t} = \mu_i \frac{\partial^2 u_i}{\partial z^2} + c\frac{\partial u_i}{\partial z} + h(u) \tag{7.1}$$

for which $u = u^*(z)$ is a stationary solution. Clearly Theorem 7.1 remains valid with an obvious redefinition of sub- and supersolutions. We shall use comparison functions of the form

$$\underline{u}_i(z,t) = u_i^*(z - z_0 + \alpha\epsilon\eta_i e^{-\sigma t}) - \epsilon\eta_i e^{-\sigma t},$$
$$\bar{u}_i(z,t) = u_i^*(z + z_1 - \alpha\epsilon\eta_i e^{-\sigma t}) + \epsilon\eta_i e^{-\sigma t},$$

with ϵ chosen so small that $\underline{u}$, $\bar{u} \in U$.

THEOREM 7.2. *For sufficiently small ϵ and σ, sufficiently large α,and any z_0, z_1, the functions $\underline{u}$, $\bar{u}$ are a sub- and supersolution, respectively, of* (7.1).

Proof. This is a generalization of [7, Lem. 4.1]. Let

$$N_i(u_i) = -\frac{\partial u_i}{\partial t} + \mu_i\frac{\partial^2 u_i}{\partial z^2} + c\frac{\partial u_i}{\partial t} + h_i(u).$$

With $q = \eta e^{-\sigma t}$, the argument of the u_i being $z - z_0 + \alpha\epsilon\eta_i e^{-\sigma t}$,

$$\begin{aligned} N_i(\underline{u}_i) &= \mu_i\frac{d^2u_i^*}{dz^2} + c\frac{du_i^*}{dz} + h_i(u^* - \epsilon q) - \sigma\epsilon q_i + \alpha\sigma\epsilon q_i\frac{du_i^*}{dz} \\ &= -h_i(u^*) + h_i(u^* - \epsilon q) - \sigma\epsilon q_i + \alpha\sigma\epsilon q_i\frac{du_i^*}{dz}. \end{aligned} \tag{7.2}$$

The above remark concerning the derivative in the direction η together with C^1 continuity shows that for small enough ϵ_0 and δ there is a constant k such that

$$h(u^* - \epsilon q) - h(u^*) \geq k\epsilon q, \qquad (0 \leq \epsilon \leq \epsilon_0,\ |B - u^*| < \delta). \tag{7.3}$$

However, $du^*/dz > 0$ as the wave is monotone, so when (7.3) holds,

$$N_i(u_i) > (k - \sigma)q_i > 0, \qquad (\sigma < k).$$

A similar argument is clearly valid for $|u^* - A| < \delta$ by choosing δ and ϵ_0 smaller if necessary.

The result will follow if we can show that a similar result holds also for u^* in neither of the above neighborhoods. Since the orbit is strictly monotone, there is a number δ' depending only on δ such that in this region $|u_i^*| > \delta'$ and $|u_i^* - 1| > \delta'$ for each i. Then, du_i^*/dz is positive and bounded away from zero. Thus from (7.2) for α large enough, $N_i(\underline{u}_i) > 0$. This proves that under the stated assumptions, $\underline{u}$ is a subsolution, and a similar proof shows that $\bar{u}$ is a supersolution. □

THEOREM 7.3. *All bistable waves have the same wave speed.*

Proof. In addition to the travelling wave u^* with speed c considered above, suppose there is a second travelling wave u^{**} with speed c_1. It is clearly possible to choose z_0 and z_1 such that

$$\underline{u}(z,0) \leq u^{**}(z,0) \leq \bar{u}(z,0).$$

Thus by Theorem 7.1, for all $t \geq 0$,

$$\underline{u}(z,t) \leq u^{**}(z,t) \leq \bar{u}(z,t).$$

However, the left and right sides approach $u^*(z - z_0)$ and $u^*(z + z_1)$, respectively, as $t \to \infty$. If now $c_1 \neq c$, then u^{**} could not be contained in this way between two translates of the stationary solution u^*. This contradiction therefore establishes the theorem. □

This result leaves open the question of whether there can be two distinct travelling waves for given wave speed. We are not able to resolve this in general, although the result can be shown to hold for two species (that is $n = 2$), by using an argument based on a comparison of the projection of the orbits on $\mathbf{R}^2$. The argument does not appear to generalize to more than two species.

THEOREM 7.4. *Every monotone bistable travelling wave is L_∞ stable.*

Proof. Let $\delta > 0$ and let $u(x,t)$ be an exact solution of (6.1) such that $|u(z,0) - u^*(z)| < \delta$ for all z. By the construction of the sub- and supersolutions, it is clear that there are numbers K and K' independent of δ such that $\epsilon < K\delta$ and

$$\underline{u}(z,0) < u(z,0) < \bar{u}(z,0),$$

$$|\underline{u}(z,t) - u^*(z)| + |\bar{u}(z,t) - u^*(z)| < K'\delta, \qquad ((z,t) \in \mathbf{R} \times \mathbf{R}_+).$$

By Theorem 7.1, $\underline{u}(z,t) \leq u(z,t) \leq \bar{u}(z,t)$, and combining these inequalities we obtain the relation

$$|u(z,t) - u^*(z)| < L'\delta, \qquad (t \geq 0).$$

This completes the proof of stability. □

Appendix. The proofs of Theorems 3.1 and 3.2 are presented here. It is assumed that the reader is familiar with the Conley index theory (see [2], [22], [23]), Conley's connection matrix (see [8]) and transition matrices (see [9], [17], [18], [21]), however, we begin by establishing some notation.

In what follows S always denotes an isolated invariant set and N an isolating neighborhood. Let (N,L) be an index pair for S, then the homotopy type of the pointed space $(N/L,[L])$ is usually referred to as the Conley index. There exists, however, a finer version of the index which will be used in the proof. First a definition. A *connected simple system* consists of a collection I_o of pointed spaces along with a collection I_m of homotopy classes of maps between these such that:

1. $\hom(X,X') = \{[f] \in [X,X'] | [f] \in I_m\}$ is nonempty and consists of a single element for each ordered pair X, X' of spaces in I_o;
2. if $X, X', X'' \in I_o$, $[f] \in \hom(X,X')$, and $[f'] \in \hom(X',X'')$, then $[f' \circ f] \in \hom(X,X'')$;
3. $\hom(X,X) = \{[1_X]\}$ for all $X \in I_m$.

Recall ([2], [20]) that the Conley index of S forms a connected simple system where $I_o = \{(N/L,[L]) | (N,L)$ is an index pair for $S\}$ and I_m consists of the flow defined maps between the elements of I_o. The connected simple system of the Conley index of S is denoted by $I(S)$. The following result [2], [22] is crucial to our analysis.

FACT. *If (A^c, A^{c*}) is an attractor repeller pair for S^c which continues for $c \in \mathbf{R}$ and $S^{c_i} = A^{c_i} \cup A^{c_i *}$ when $i = 0,1$, but $I(S^{c_0}) \not\sim I(S^{c_1})$, then for some $c \in (c_0, c_1)$ there exists a connecting orbit from A^{c*} to A^c.*

The $\mathbf{Z}_2$ homology Conley index of S is denoted and defined by

$$CH_*(S) := H_*(N/L,[L];\mathbf{Z}_2).$$

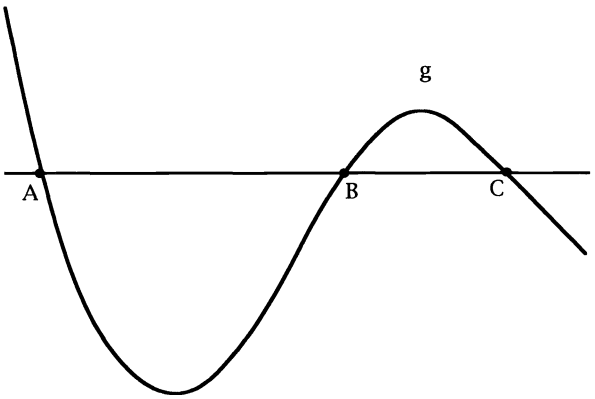


FIG. A.1

Given a Morse decomposition of S, $\mathbf{M}(S) = \{M(i) | i \in (P, >)\}$, where P is the indexing set for the Morse sets $M(i)$ and $>$ is a strict partial order on P, recall that the connection matrix

$$\Delta : \bigoplus_{i \in P} CH_*(M(i)) \longrightarrow \bigoplus_{i \in P} CH_*(M(i)).$$

We shall always take the direct sum according to a nonincreasing order in P. This has the effect of making Δ into a strictly upper triangular matrix.

The proof of Theorem 3.1 breaks down into three steps. The first is to study the dynamics of the two-dimensional system defined over L. The second is to relate these results with the the full travelling wave system ($\mathrm{TW}^{c,1}$), and the final step is to homotope these results back to the complicated system ($\mathrm{TW}^{c,0}$). As will become obvious, this theorem is a triviality since most of the difficulties have been assumed away.

Since L is an invariant line under the reaction system (R^1) we can think of the dynamics on L as being given by a scalar differential equation

$$\xi' = g(\xi), \tag{A.1}$$

where $\xi \in \mathbf{R}$ and $g : \mathbf{R} \to \mathbf{R}$. Assumption I implies that $g(\xi)$ has exactly 3 zeros, $\{a, b, c\}$, and must have the form given in Fig. A.1. Now define $G : \mathbf{R} \to \mathbf{R}$ by $G'(\xi) = g(\xi)$. Obviously $G(a) > G(c)$ and $G(b) > G(c)$. Without loss of generality assume that $G(a) > G(b)$. The corresponding travelling wave system can then be written as

$$\begin{aligned} \xi' &= \eta, \\ \eta' &= -c\eta - G'(\xi), \end{aligned} \tag{A.2c}$$

where $c \in \mathbf{R}$ is the wave speed and $\eta \in \mathbf{R}$. Clearly the fixed points of (A.2c) are $\{\alpha = (a, 0), \beta = (b, 0), \gamma = (c, 0)\}$. Now define

$$V(\xi, \eta) := \frac{\eta^2}{2} + G(\xi),$$

then $dV/dt = -c\eta^2$, i.e., for $c > 0$, V is a Lyapunov function for (A.2). Let $s(c)$ denote the set of bounded solutions to (A.2c), then by Assumption I, $s(c)$ is an isolated invariant set.

LEMMA A.1. *Let $c > 0$, then:*

(i) $\mathbf{M}(s(c)) = \{\alpha, \beta, \gamma | \alpha > \beta > \gamma\}$ *is a Morse decomposition of* $s(c)$.

(ii) $CH_k(s(c)) \approx CH_k(\alpha) \approx CH_k(\beta) \approx \begin{cases} \mathbf{Z}_2, & \text{if } k = 1; \\ 0, & \text{otherwise.} \end{cases}$

(iii) $CH_k(\gamma) \approx \begin{cases} \mathbf{Z}_2, & \text{if } k = 0; \\ 0, & \text{otherwise.} \end{cases}$

(iv) *There exists* c_l *such that for all* $c \in (0, c_l)$ *the connection matrix takes the form*

$$\Delta_l = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

(v) *There exists* c_r *such that for all* $c \in (0, c_r)$ *the connection matrix takes the form*

$$\Delta_r = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Proof. (i) follows from $dV/dt = -c\eta^2$ and the assumption that $G(\alpha) > G(\beta) > G(\gamma)$. (ii) and (iii) follow from the fact that the unstable manifolds of α and β are one-dimensional, while γ is an attracting critical point. (iv) and (v) follow from a simple phase portrait analysis. □

Since L is an attracting invariant line under (R^1) one immediately obtains similar results for S^c, the set of bounded solutions for $(\mathrm{TW}^{c,1})$, in a neighborhood of $s(c)$ as a subset of $\mathbf{R}^{2n}$.

LEMMA A.2. *Let $c > 0$, then:*

(i) $\mathbf{M}(S^c) = \{A_0^1, B_0^1, C_0^1 | A_0^1 > B_0^1 > C_0^1\}$ *is a Morse decomposition of* S^c.

(ii) $CH_k(S^c) \approx CH_k(A_0^1) \approx CH_k(B_0^1) \approx \begin{cases} \mathbf{Z}_2, & \text{if } k = 1; \\ 0, & \text{otherwise.} \end{cases}$

(iii) $CH_k(C_0^1) \approx \begin{cases} \mathbf{Z}_2, & \text{if } k = 0; \\ 0, & \text{otherwise.} \end{cases}$

(iv) *There exists* c_l *such that for all* $c \in (0, c_l]$ *the connection matrix takes the form defined above.*

(v) *There exists* c_r *such that for all* $c \in [c_r, \infty)$ *the connection matrix takes the form defined above.*

LEMMA A.3. *The transition matrix between* Δ_l *and* Δ_r *is*

$$T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Proof. Since $A_0^1 > B_0^1 > C_0^1$ is an admissible ordering which continues over the parameter range $c \in (0, \infty)$, T must be an upper triangular matrix with respect to this ordering. Furthermore, T is a degree 0 map of the form $I + N$ where I is the identity and N is strictly upper triangular [18]. Finally, T satisfies $T\Delta_l = \Delta_r T$. The only matrix which meets all these conditions is the desired matrix. □

COROLLARY A.4. *There exists* $c \in [c_l, c_r]$ *for which an* $A^1 \to B^1$ *travelling wave occurs.*

Let $I((N^{c,1})^\infty)$ denote the connected simple system of the Conley index for $(N^{c,1})^\infty$ under the flow induced by $(\mathrm{TW}^{c,1})$.

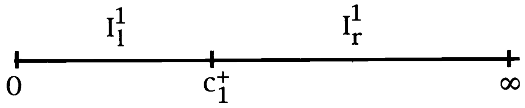


FIG. A.2

COROLLARY A.5 [16, §6]. *For $c \in (0, c_l]$, $I((N^{c,1})^\infty) = I_l^1$ and for $c \in [c_r, \infty)$, $I((N^{c,1})^\infty) = I_r^1$. Furthermore, $I_r^1 \not\succ I_l^1$.*

It might be helpful to visualize the content of this last corollary. Let $c_1^+ := \inf\{c_r\}$, and $c_1^- := \sup\{c_l\}$. Of course $c_1^- \leq c_1^+$. To simplify the picture assume that $c_1^- = c_1^+$. Then Corollary 5 is summarized in Fig. A.2.

Recall that $dV/dt = -c\eta^2$ and that $G(a) > G(b)$. Thus it is impossible for an $A_0^1 \to B_0^1$ to exist if $c \leq 0$. Therefore, for the isolated invariant set N_1, the connected simple system I_l^1 continues over $c \leq 0$, i.e., Fig. A.2 extends to Fig. A.3.

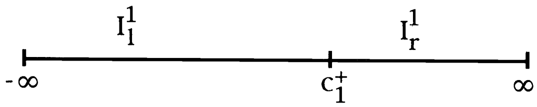


FIG. A.3

Now consider the homotopy from $\sigma = 1$ back to the original system at $\sigma = 0$. First, by [3] there exists an a priori bound Γ such that $|c| > \Gamma$ implies that there does not exist an $A_0^\sigma \to B_0^\sigma$ solution to $(\mathrm{TW}^{c,\sigma})$ for all $\sigma \in [0,1]$. Therefore, $I_l^\sigma \not\succ I_r^\sigma$ where the connected simple systems are computed for $-c < -\Gamma$ and $c > \Gamma$, respectively. In particular, $I_l^0 \not\succ I_r^0$ and therefore an $A_0 \to B_0$ solution exists. This completes the proof of Theorem 3.1. For a pictorial description of this argument see Fig. A.4.

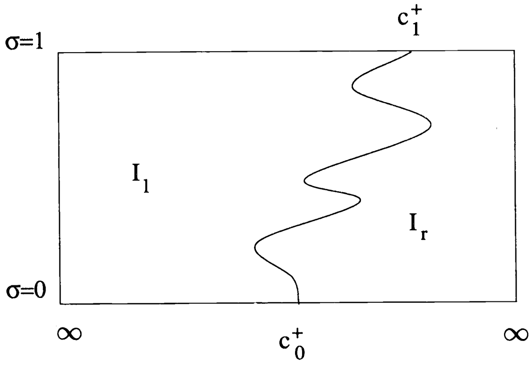


FIG. A.4

Proof of Theorem 3.2. Let $S = A \cup C \cup C(C, A)$ which is obviously an isolated invariant set. Let N be an isolating neighborhood of S under the reaction flow. Since A is an attractor, the stable manifold of A intersects transversely with the unstable manifold of C. By a result of [16], this implies that $CH_k(S) \approx 0$ for all $k = 0, 1, \ldots$. Now by [3], there exists $\hat{c} > 0$ such that given $c < -\hat{c} \ll -1$, there exists an isolating neighborhood $\bar{N}$ in the travelling wave system such that $(A_0 \cup C_0) \subset \bar{N}$ and $CH_{k+n}(\bar{N}^\infty) \approx CH_k(S) \approx 0$ for all $k = 0, 1, \ldots$. Of course the indices of A_0 and C_0 are not trivial, and hence there exists a connecting orbit from C_0 to A_0. Now, by a symmetry argument, one has that for $c > \hat{c}$, there exists an A_0 to C_0 connecting orbit. □

A final remark is necessary. The introduction of connection matrices and transition matrices in the proof of Theorem 3.1 is not really necessary. Since (A.2c) is two-dimensional, a phase portrait analysis is sufficient to convince oneself that the connected simple systems are different at high and at low wave speeds. On the other hand, it is clear that for some problems Assumption I is too restrictive (see [19]). A weaker assumption would be the following:

Assumption I*. For the reaction system (R_1) there exists a k-dimensional attracting invariant subspace $L \subset \mathbf{R}^n$. Furthermore, the dynamics on L under (R_1) can be described by a gradient flow, i.e.,

$$\xi' = \nabla G(\xi).$$

In this case, under this assumption if one could determine the connection matrices for $c \approx \infty$ and for $c \approx 0$ and they were related by an appropriate transition matrix, then the proof, and hence the result of Theorem 3.1 is still valid.

Acknowledgments. Apart from the acknowledgment already made to Paul Fife in §7, V. Hutson wishes particularly to thank him for sharing his exceptional insights on travelling wave problems. V. Hutson wishes also to acknowledge generous support from the Leverhulme Trust, which made possible a visit to the Georgia Institute of Technology in 1989, and also support from Sheffield University and Georgia Tech for a short visit in 1990.

REFERENCES

[1] D. Aronson and H. Weinberger, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. Math., 30 (1978), pp. 33–76.

[2] C. C. Conley, *Isolated Invariant Sets and the Morse Index*, CBMS Lecture Notes 38, American Mathematical Society, Providence, RI, 1978.

[3] C. C. Conley and P. Fife, *Critical manifolds, travelling waves and an example from population genetics*, J. Math. Bio., 14 (1982), pp. 159–176.

[4] C. C. Conley and R. Gardner, *An application of the generalized Morse index to travelling wave solutions of a competitive reaction diffusion model*, Indiana Univ. Math. J., 33 (1989), pp. 319–343.

[5] S. Dunbar, *Travelling wave solutions of diffusive Lotka–Volterra equations: a heteroclinic connection in* $\mathbf{R}^4$, Trans. Amer. Math. Soc., 286 (1984), pp. 557–594.

[6] P. C. Fife, *Mathematical Aspects of Reacting and Diffusing Systems*, Lecture Notes in Biomath.. 28, Springer-Verlag, New York, 1979.

[7] P. C. Fife and J. B. McCleod, *The approach of solutions of nonlinear diffusion equations to travelling front solutions*, Arch. Rational Mech. Anal., 65 (1977), pp. 335–361.

[8] R. Franzosa, *The connection matrix theory for Morse decompositions*, Trans. Amer. Math. Soc., 311 (1989), pp. 781–803.

[9] R. Franzosa and K. Mischaikow, *Algebraic Transition Matrices*, in preparation.

[10] R. Gardner, *Existence and stability of travelling wave solutions of competition models: a degree theoretical approach*, J. Differential Equations, 44 (1982), pp. 363–364.

[11] ———, *Existence of travelling wave solution of predator-prey systems via the Conley index*, SIAM J. Appl. Math., 44 (1984), pp. 56–76.

[12] M. Hirsch, *Systems of differential equations which are competitive or cooperative.* I: *Limit sets*, SIAM J. Math. Anal., 13 (1982), pp. 167–179.

[13] V. Hutson, *Stability in a reaction diffusion model of mutualism*, SIAM J. Math. Anal., 17 (1986), pp. 58–66.

[14] V. Hutson, R. Law, and D. Lewis, *Dynamics of ecologically obligate mutualisms-effects of spatial diffusion on resilience of the interacting species*, American Naturalist, 126 (1985), pp. 465–469.

[15] G. Kassen and W. Troy, *The stability of traveling wave front solutions of a reaction-diffusion system*, SIAM J. Appl. Math., 41 (1981), pp. 145–167.

[16] C. McCord, *The connection map for attractor-repeller pairs*, Trans. Amer. Math. Soc., 307 (1988), pp. 195–203.

[17] C. McCord and K. Mischaikow, *Connected Simple Systems, Transition Matrices, and Heteroclinic Bifurcations*, Trans. Amer. Math. Soc., to appear.

[18] K. Mischaikow, *Transition Systems*, Proc. Roy. Soc. Edinburgh, 112A (1989), pp. 155–175.

[19] ———, *Travelling waves for a cooperative and a competitive-cooperative system*, in Viscous Profiles and Numerical Methods for Shock Waves, M. Shearer, ed., Society for Industrial and Applied Mathematics, Philadelphia, 1991, pp. 125–141.

[20] K. Mischaikow and J. Reineck, *Travelling waves in predator-prey systems*, SIAM J. Math. Anal., 24 (1993), to appear.

[21] J. Reineck, *Connecting orbits in one-parameter families of flows*, Ergodic Theory Dynamical Systems, 8* (1988), pp. 359–374.

[22] D. Salamon, *Connected simple systems and the Conley index of isolated invariant sets*, Trans. Amer. Math. Soc., 291 (1985), pp. 1–41.

[23] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

[24] W. Walter, *Differential und Integral Ungleichungen*, Springer-Verlag, Berlin, 1964.

SIAM J. MATH. ANAL.
Vol. 24, No. 4, pp. 1009-1029, July 1993

009

STABILITY OF EQUILIBRIA FOR A CLASS OF TIME-REVERSIBLE, $D_n \times O(2)$-SYMMETRIC HOMOGENEOUS VECTOR FIELDS*

I-HENG McCOMB†‡ AND CHJAN C. LIM†§

Abstract. First-order, time-reversible n-body problems in three-space whose velocity fields consist of sums of identical two-body interactions are studied under a set of natural symmetry assumptions. Up to linearization about maximally symmetric equilibria, the entire class is shown to be represented by a two-parameter normal form. The symmetries of the class are used to find formulas for the eigenvalues of the linearized problems. The class of problems is divided into two families, one in which vector field components in the spatial directions act in concert, and one in which they act in opposition. When the components act in concert, the equilibria are (i) unstable when interaction strength grows with distance, (ii) stable when interaction strength decays and $n = 3, 4$, and (iii) stable or unstable when interaction strength decays and $n > 4$, depending as the singularity of the vector field varies across a critical value. When components act in opposition, stability and instability are interchanged. A nonlinear application of this analysis is the establishment of symmetric near-equilibrium periodic solutions.

Key words. time reversibility, symmetry, normal modes, stability

AMS subject classifications. 34, 70K15

Introduction. We consider the first-order three-dimensional motion of n bodies under a time-reversible vector field given by the sum of identical two-body interactions, where the time reversal reverses one spatial direction while fixing the other two. We impose a set of natural symmetry assumptions, and ask what can be said about the linearizations of these vector fields about maximally symmetric equilibria *entirely on the basis of the symmetries.*

First-order n-body problems have $3n$-dimensional phase space (whereas second-order problems such as Newtonian mechanical n-body problems have $6n$-dimensional phase space). In the notation we will use we have vector fields

$$\frac{dr}{dt} = f(r), \tag{0.1}$$

where $r = (r_1, \ldots, r_n)$ with $r_j = (x_j, y_j, z_j) \in \mathbf{R}^3$ and $f = (f_1, \ldots, f_n)$ with $f_j = (f_j^x, f_j^y, f_j^z)$. Since the vector field consists of the sum of identical two-body interactions (and under translational symmetry, of course), $f(r)$ is of the form

$$f_j(r) = \sum_{k \neq j} \frac{F(r_j - r_k)}{|r_j - r_k|^\alpha} + \text{const}, \tag{0.2}$$

where $F = (F^x, F^y, F^z)$ and the constant is independent of j.

* Received by the editors April 6, 1992; accepted for publication (in revised form) October 30, 1992.

† Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York 12180-3590.

‡ This author is supported by a Department of Education Fellowship.

§ This author is funded by National Science Foundation grant DMS-9006091.

Time reversibility is a symmetry. A vector field problem is said to be time-reversible when there exists a phase space reflection R which anticommutes with the vector field. For a time-reversible vector field, whenever $r(t)$ is a solution, so is $Rr(-t)$. A time-symmetric solution is one for which $r(t)=Rr(-t)$, or equivalently, any solution which passes through the time-symmetry plane Fix R. More general reversible systems are discussed in detail in Sevryuk [S]. Here we assume that $f(r)$ is time-reversible with respect to reflection across the (x, y)-plane, i.e.,

$$\text{(H1)} \qquad Rf(r)=-f(Rr),$$

where

$$(0.3) \qquad Rr=(\bar{r}_1, \ldots, \bar{r}_n), \qquad \bar{r}_j=(x_j, y_j, -z_j).$$

It is natural to assume that the time-reversal direction (here the z-direction) is the only distinguished direction. Hence we assume $f(r)$ is $O(2)$-symmetric with respect to the z-axis. Letting R_θ denote rotation by θ about the origin applied to each of the n pairs (x_j, y_j), and K_θ denote reflection in the (x, y)-plane across the line obtained when the x-axis is rotated by θ about the origin, we have

$$\text{(H2)} \qquad R_\theta f(r)=f(R_\theta r), \quad K_\theta f(r)=f(K_\theta r) \quad \forall \theta.$$

Finally, we assume that $f(r)$ is homogeneous. To cast this assumption as a symmetry of the problem, we write: there exists β such that

$$\text{(H3)} \qquad f(\delta r)=\delta^{-\beta} f(r)$$

for all positive real δ (where equality holds up to a constant corresponding to pure translation). We will refer to β as the "singularity" of the vector field. Homogeneity is a natural assumption if we consider the approximation of a problem by its leading term.

We show that vector field problems of the form (0.1), (0.2) are D_n-symmetric. Under D_n symmetry and time reversibility (H1), regular horizontal n-gons are the only maximally symmetric equilibria. In particular, we consider the equilibria given by r^0, where

$$(0.4) \qquad r_j^0=\left(\cos\frac{2\pi j}{n}, \sin\frac{2\pi j}{n}, 0\right).$$

Note that these equilibria are actually relative equilibria in the sense that

$$(0.5) \qquad f_j(r^0)=\text{const} \quad (\text{independent of } j) \quad \forall j.$$

Any translation, rotation, or scaling of these equilibria are, of course, also equilibria (corresponding to translational, $O(2)$, and homogeneity symmetries of the problem).

When vertical and horizontal components of the vector field act in concert, we obtain the following stability results:

(i) When interaction strength grows over distance, regular horizontal n-gons are unstable;

(iia) For $n=3, 4$, when interaction strength decays over distance, regular horizontal n-gons are stable;

(iib) For $n>4$, when interaction strength decays over distance, regular horizontal n-gons are stable or unstable, depending as the singularity β passes through a "stability threshold" $\beta^*(n)<\infty$, where stability refers to ellipticity. (Hyperbolic stability is disallowed by time reversibility, so ellipticity is the strongest stability obtainable by linear analysis.) Numerical studies indicate that for problems of type (iib), $\beta^*(n)$

decreases with increasing n, i.e., that less singular cases are more stable for decaying interaction. When the components of the vector field act in opposition, stability and instability are of course interchanged.

To obtain these results, we use the symmetry properties of the class of problems to exhibit a normal form for linear analysis of the entire class. We show that up to linearization about regular horizontal n-gons, our class of vector field problems reduces to two 2-parameter families, parametrized by the relative magnitudes of the vertical and horizontal components, and by the singularity strength β. One physical example which may be treated under this setting is the sedimentation under gravity of n small clustered spheres in a highly viscous fluid, studied experimentally by [JMS], and analytically under the Stokeslet model (infinite viscosity and point particle approximation) in [H], [CLLS], [GKL], and [TK]. The Stokeslet interaction is given by

$$F(x, y, z)=\left(\frac{xz}{(x^2+y^2+z^2)^{3/2}}, \frac{yz}{(x^2+y^2+z^2)^{3/2}}, \frac{x^2+y^2+2z^2}{(x^2+y^2+z^2)^{3/2}}\right).$$

We obtain formulas for the eigenvalues, essentially as in [H], but for the whole class, and in terms of the assumed symmetries. The (spectral) stability of the equilibria is studied analytically and numerically. The Stokeslet problem falls under case (ii) above, with $\beta=1$. We have $\beta^*(7)<1<\beta^*(6)$ so that regular horizontal n-gons are stable for $n=3, 4, 5, 6$ and unstable for $n \geqq 7$, in agreement with [JMS] and [H]. Another example is the dipole-type interaction given by

$$F(x, y, z)=\left(\frac{3xz}{(x^2+y^2+z^2)^{5/2}}, \frac{3yz}{(x^2+y^2+z^2)^{5/2}}, \frac{-(x^2+y^2)+2z^2}{(x^2+y^2+z^2)^{5/2}}\right).$$

Here vertical and horizontal components act in opposition and $\beta=3$.

The linear analysis of this class of problems, besides providing stability information, serves as a basis for nonlinear results and investigations. For example, for $0<\beta<\beta^*(n)$, a time-reversible Lyapunov center type theorem for spatially symmetric systems [GKL] applies, and we are able to establish the existence of families of symmetric, near-equilibrium periodic solutions when a nonresonance condition holds.

1. D_n symmetry. We have stated as assumptions the time reversibility and $O(2)$ symmetry of our class. We now show that D_n symmetry is a consequence of the identical two-body interaction form of $f(r)$ given by (0.2). Let

$$\begin{aligned}\mathscr{C}(r_1, r_2, \ldots, r_n)&=(r_2, \ldots, r_n, r_1),\\ \mathscr{T}(r_1, r_2, \ldots, r_n)&=(r_n, \ldots, r_2, r_1),\end{aligned} \tag{1.1}$$

and

$$\hat{\mathscr{C}}=R_{-2\pi/n}\mathscr{C}, \tag{1.2a}$$

$$\hat{\mathscr{T}}=K_{\pi/n}\mathscr{T}, \tag{1.2b}$$

with R_θ and K_θ as defined in (H2). From the form of $f(r)$ given by (0.2), we see that

$$\mathscr{C}f(r)=f(\mathscr{C}r) \quad \text{and} \quad \mathscr{T}f(r)=f(\mathscr{T}r). \tag{1.3}$$

By $O(2)$ symmetry (H2), we have

$$\hat{\mathscr{C}}f(r)=f(\hat{\mathscr{C}}r) \quad \text{and} \quad \hat{\mathscr{T}}f(r)=f(\hat{\mathscr{T}}r). \tag{1.4}$$

Moreover,

$$\hat{\mathscr{C}}r^0=\hat{\mathscr{T}}r^0=r^0. \tag{1.5}$$

Hence we find

$$\gamma f(r) = f(\gamma r) \quad \gamma r^0 = \cdot\cdot^0 \quad \forall \gamma \in \Gamma = (\hat{\mathscr{C}}, \hat{\mathscr{T}}) \cong D_n. \tag{1.6}$$

2. Isotypic decomposition. In the eigenvalue analysis to follow, we will require the Γ-isotypic decomposition for our space $\mathbf{R}^{3n}$ (cf. [GSS]). Recall that each Γ-isotypic component of a space is the direct sum of all Γ-isomorphic copies occurring in a decomposition of the space into Γ-irreducible subspaces (i.e., into Γ-invariant subspaces with no proper, nontrivial Γ-invariant subspaces). Every space has a unique Γ-isotypic decomposition. We begin by listing the one- and two-dimensional irreducible representations of D_n (up to isomorphism). When n is odd, the one-dimensional irreducible representations are

$$W_{++}, \quad \hat{\mathscr{C}} = 1, \quad \hat{\mathscr{T}} = 1, \tag{2.1a}$$

$$W_{+-}, \quad \hat{\mathscr{C}} = 1, \quad \hat{\mathscr{T}} = -1. \tag{2.1b}$$

When n is even, there are in addition

$$W_{-+}, \quad \hat{\mathscr{C}} = -1, \quad \hat{\mathscr{T}} = 1, \tag{2.1c}$$

$$W_{--}, \quad \hat{\mathscr{C}} = -1, \quad \hat{\mathscr{T}} = -1. \tag{2.1d}$$

There are int $((n-1)/2)$ distinct two-dimensional irreducible representations:

$$W_k \cong \mathbf{C}, \qquad \hat{\mathscr{C}}_z = e^{i2\pi k/n} z, \quad \hat{\mathscr{T}} z = \bar{z}, \tag{2.2}$$

where $z \in \mathbf{C}$ and $(\bar{\ })$ denotes complex conjugation.

We proceed with the isotypic decomposition of $\mathbf{R}^{3n}$. We write

$$\mathbf{R}^{3n} = H \oplus V, \tag{2.3}$$

where

$$H = \{ r \in \mathbf{R}^{3n} \mid z_j = 0 \quad \forall j \} \tag{2.4a}$$

and

$$V = \{ r \in \mathbf{R}^{3n} \mid x_j = y_j = 0 \quad \forall j \}. \tag{2.4b}$$

We note that H and V are Γ-invariant, i.e., $\Gamma H \subseteq H$ and $\Gamma V \subseteq V$. We decompose H as

$$H = H_0 \oplus H_1 \oplus \cdots \oplus H_{n-1}, \tag{2.5}$$

where

$$H_l = \{ r \in H \mid r_{j+1} = R_{2\pi l/n} r_j \}. \tag{2.6}$$

Identifying $r \in H_l$ such that $r_1 = \rho$ with $\rho = (\xi, \eta, 0) \in \mathbf{R}^3$, we have

$$\hat{\mathscr{C}}_\rho = R_{2\pi(l-1)/n} \rho, \qquad \hat{\mathscr{T}} \rho = K_{\pi/n} R_{-2\pi l/n} \rho. \tag{2.7}$$

We see that

$$\begin{aligned}
&H_0 \cong_\Gamma W_1, \\
&H_1 \cong_\Gamma W_{++} \oplus W_{+-}, \\
&H_l \cong_\Gamma W_{l-1} \quad \text{for } l = 2, \ldots, \text{int}\left(\frac{n-1}{2}\right) + 1, \\
&H_l \cong_\Gamma W_{n-(l-1)} \quad \text{for } l = \text{int}\left(\frac{n+2}{2}\right) + 1, \ldots, n-1, \\
&H_{(n/2)+1} \cong_\Gamma W_{-+} \oplus W_{--} \quad \text{when } n \text{ even},
\end{aligned} \tag{2.8}$$

where $\cong_\Gamma$ denotes Γ-isomorphism. Next we decompose V as

$$V = V_0 \oplus V_1 \oplus \cdots \oplus V_{\mathrm{int}\,(n/2)}, \tag{2.9}$$

where V_k is the real space

$$V_k = \mathbf{R}(v_k^r, v_k^i) \tag{2.10a}$$

with

$$\begin{aligned} v_k^r &= \left(1, \cos\frac{2\pi k}{n}, \ldots, \cos\frac{2\pi k(n-1)}{n}\right), \\ v_k^i &= \left(0, \sin\frac{2\pi k}{n}, \ldots, \sin\frac{2\pi k(n-1)}{n}\right). \end{aligned} \tag{2.10b}$$

In the basis $\{v_k^r, v_k^i\}$ for V_k, we have for any $v_k \in V_k$,

$$\begin{aligned} \hat{\mathscr{C}} v_k &= \begin{pmatrix} \cos\frac{2\pi k}{n} & \sin\frac{2\pi k}{n} \\ -\sin\frac{2\pi k}{n} & \cos\frac{2\pi k}{n} \end{pmatrix} v_k, \\ \hat{\mathscr{T}} v_k &= \begin{pmatrix} \cos\frac{2\pi k}{n} & -\sin\frac{2\pi k}{n} \\ -\sin\frac{2\pi k}{n} & -\cos\frac{2\pi k}{n} \end{pmatrix} v_k. \end{aligned} \tag{2.11}$$

We see that

$$\begin{aligned} &V_0 \cong_\Gamma W_{++}, \\ &V_k \cong_\Gamma W_k \quad \text{for } k = 1, \ldots, \mathrm{int}\left(\frac{n-1}{2}\right), \\ &V_{n/2} \cong_\Gamma W_{--} \quad \text{when } n \text{ even}. \end{aligned} \tag{2.12}$$

We arrive at the isotypic decomposition of $\mathbf{R}^{3n}$ that we want:

$$\begin{aligned} &\mathbf{R}^{3n} \cong_\Gamma W_{++}^2 \oplus W_{+-} \oplus W_1^3 \oplus \cdots \oplus W_{(n-1)/2}^3 \quad (n \text{ odd}), \\ &\mathbf{R}^{3n} \cong_\Gamma W_{++}^2 \oplus W_{+-} \oplus W_{-+} \oplus W_{--}^2 \oplus W_1^3 \oplus \cdots \oplus W_{(n-2)/2}^3 \quad (n \text{ even}). \end{aligned} \tag{2.13}$$

3. Eigenvalue structure. On the basis of the isotypic decomposition (2.13) and time reversibility (H1), we can make qualitative statements about the eigenvalues of the linearization L of f at r^0. We have

$$L \equiv df|_{r^0} = \begin{pmatrix} \partial_{r_1} f_1|_{r^0} & \partial_{r_2} f_1|_{r^0} & \cdots & \partial_{r_n} f_1|_{r^0} \\ \partial_{r_1} f_2|_{r^0} & \partial_{r_2} f_2|_{r^0} & \cdots & \partial_{r_n} f_2|_{r^0} \\ \vdots & \vdots & \ddots & \vdots \\ \partial_{r_1} f_n|_{r^0} & \partial_{r_2} f_n|_{r^0} & \cdots & \partial_{r_n} f_n|_{r^0} \end{pmatrix} = (\partial_{r_j} f_i|_{r^0}), \tag{3.1}$$

where the notation on the right means that the (i, j)th 3×3 block of L is given by

$$\partial_{r_j} f_i|_{r^0} = \begin{pmatrix} \partial f_i^x/\partial x_j|_{r^0} & \partial f_i^x/\partial y_j|_{r^0} & \partial f_i^x/\partial z_j|_{r^0} \\ \partial f_i^y/\partial x_j|_{r^0} & \partial f_i^y/\partial y_j|_{r^0} & \partial f_i^y/\partial z_j|_{r^0} \\ \partial f_i^z/\partial x_j|_{r^0} & \partial f_i^z/\partial y_j|_{r^0} & \partial f_i^z/\partial z_j|_{r^0} \end{pmatrix}. \tag{3.2}$$

Consider W any L-invariant subspace of $\mathbf{R}^{3n}$. On W, we have $LR = -RL$ by time reversibility (H1) so when $Lu = \lambda u$,

$$L(Ru) = -R(Lu) = -\lambda(Ru), \tag{3.3}$$

i.e., λ is an eigenvalue of $L|_W$ if and only if $-\lambda$ is. Now since f depends only on the relative positions $r_j - r_k$ $(j \neq k)$, H_0 and V_0 are clear!y L-invariant, and moreover, $L|_{H_0} = 0$ and $L|_{V_0} = 0$ (corresponding, respectively, to horizontal and vertical translational motion). Next, we observe that isotypic components are L-invariant. To see this, let W^s be any isotypic component. Ker L is Γ-invariant since if $u \in \operatorname{Ker} L$, we have $L(\gamma u) = \gamma(Lu) = 0$. Then $W \cap \operatorname{Ker} L$ is a Γ-invariant subspace of W and since W is irreducible, either $W \cap \operatorname{Ker} L = W$ or $W \cap \operatorname{Ker} L = 0$. We see that either $LW = 0$ or LW is isomorphic to W. Hence $LW^s \subseteq W^s$. Using the isotypic decomposition (2.13), we are now in a position to perform our eigenvalue structure analysis.

We see immediately that $L = 0$ when restricted to W^2_{++} ($L = 0$ when restricted to $V_0 \cong_\Gamma W_{++}$), W_{+-}, and W_{-+} (n even) since λ, an eigenvalue of any L-invariant subspace, implies that $-\lambda$ is also an eigenvalue. Together with the two zero eigenvalues corresponding to horizontal translation, we have so far counted five zero eigenvalues when n is odd, and an additional zero eigenvalue when n is even.

Next we consider the components W^2_{--} and W^3_k. W_{--} is one-dimensional so it is trivially absolutely irreducible. The action of Γ on $W_k \cong \mathbf{C}$ (given by (2.2)) is also absolutely irreducible (in the sense that when W_k is viewed as a real space, the only linear maps that commute with Γ are multiples of the identity). Hence in an appropriate basis,

$$L|_{W^s} = \begin{pmatrix} a_{11}I & \cdots & a_{1s}I \\ \vdots & \ddots & \vdots \\ a_{s1}I & \cdots & a_{ss}I \end{pmatrix}, \tag{3.4}$$

where $I = 1, s = 2$ for $W^s = W^2_{--}$ and $I = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, $s = 3$ for $W^s = W^3_k$. We see that the eigenvalues of $L|_{W^s}$ are exactly the eigenvalues of

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1s} \\ \vdots & \ddots & \vdots \\ a_{s1} & \cdots & a_{ss} \end{pmatrix}, \tag{3.5}$$

each with multiplicity dim W. Since λ is an eigenvalue of A if and only if $-\lambda$ is, we see that W^2_{--} has eigenvalues $\pm\lambda$, while W^3_k has eigenvalues $0, \pm\lambda$ each of multiplicity two. Since there are int $((n-1)/2)$ components W^3_k, we have counted $2 \cdot$ int $((n-1)/2)$ zero eigenvalues and int $((n-1)/2)$ pairs $\pm\lambda$ of multiplicity two. When n is even, W^2_{--} contributes an additional pair $\pm\lambda$. However, the two zero eigenvalues from W^3_1 correspond to H_0 so they have already been counted. We actually have $2 \cdot$ int $((n-3)/2)$ additional zero eigenvalues.

In fact, we can say more. Noting that W^2_{--} arises (in the n even case) from $V_{n/2}$ and part of $H_{n/2+1}$, we see that in some basis, $R|_{W^2_{--}} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$. Then if in this basis $A = (a_{ij})$, time reversibility (H1) implies

$$\begin{pmatrix} a_{11} & a_{12} \\ -a_{21} & -a_{22} \end{pmatrix} = \begin{pmatrix} -a_{11} & a_{12} \\ -a_{21} & a_{22} \end{pmatrix}, \tag{3.6}$$

i.e., $a_{11} = a_{22} = 0$. We see that the eigenvalues of A are real or pure imaginary. In the case of W^3_k, we have (in some basis)

$$R|_{W^s_k} = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & -I \end{pmatrix}, \tag{3.7}$$

where $I = \binom{1\ 0}{0\ 1}$. If in this basis, $A = (a_{ij})$, $RL = -LR$ (H1) implies

$$(3.8) \qquad \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ -a_{31} & -a_{32} & -a_{33} \end{pmatrix} = \begin{pmatrix} -a_{11} & -a_{12} & a_{13} \\ -a_{21} & -a_{22} & a_{23} \\ -a_{31} & -a_{32} & a_{33} \end{pmatrix},$$

i.e., $a_{11} = a_{12} = a_{21} = a_{22} = a_{33} = 0$. Hence the eigenvalues of A satisfy $\lambda(\lambda^2 - a_{31}a_{13} - a_{32}a_{23}) = 0$. Again, all eigenvalues are real or pure imaginary.

In summary, we have the following.

THEOREM I (eigenvalue structure). *Consider $dr/dt = f(r)$, where $f(r)$ has the form* (0.2) *and satisfies the symmetry relations* (H1)-(H3). *Then the linearization of f at maximally symmetric equilibria has the following*:

(i) 5 *zero eigenvalues corresponding to three dimensions of translational motion, one dimension of scale contraction or expansion, and one dimension of rotation about center (average position)*;

(ii) 1 *additional zero eigenvalue when n is even, corresponding to a type of horizontal displacement which produces vertical translation only—the relevant subspace is that spanned by $r \in H_{n/2+1}$, where $r_1 = (-\sin(2\pi/n), \cos(2\pi/n), 0)$*;

(iii) int $((n-1)/2)$ *pairs of real or pure imaginary eigenvalues $\pm\lambda$, each of multiplicity two*;

(iv) $2 \cdot \operatorname{int}((n-3)/2)$ *more zero eigenvalues (forced by time reversibility)*;

(v) 1 *additional pair of real or pure imaginary eigenvalues $\pm\lambda$ when n is even.*

4. Relation between eigenvalues of L and L^2. It turns out that it is convenient to obtain the eigenvalues of L from the eigenvalues of $L^2|_V$, so we exhibit the relation between the two sets of eigenvalues. By time reversibility (H1) we have $fR = -Rf$ so that $L^2R = RL^2$. We see that $L^2H \subseteq H$ and $L^2V \subseteq V$.

Consider (λ, u) an eigenvalue/vector pair for L. From (3.3), $Lu = \lambda u$ and $L(Ru) = -\lambda(Ru)$, so for $\lambda \neq 0$, u must have nonzero vertical part. Writing $u = h + v$ where $h \in H$, $v \in V$, we have $L^2h + L^2v = \lambda^2h + \lambda^2v$. Since H and V are L^2-invariant, we have in particular,

$$(4.1) \qquad L^2v = \lambda^2v.$$

Hence every nonzero eigenvalue of L is the square root of an eigenvalue of $L^2|_V$.

Conversely, let (μ, v) be an eigenvalue/vector pair for $L^2|_V$. Since $LW^s \subseteq W^s$ for any isotypic component W^s, each W^s is also L^2-invariant. For some W^s, there exists $u \in W^s$ such that $L^2u = \mu u$. For $L|_{W^s} \neq 0$, we know $L|_{W^s}$ has exactly one pair of eigenvalues $\pm\lambda$ (of some multiplicity but semisimple) and the rest zero. Letting $E_\pm$ denote the eigenspaces corresponding to $\pm\lambda$, we can write $u = u_+ + u_- + n$ such that $u_+ \in E_+$, $u_- \in E_-$, and $n \in \operatorname{Ker} L$. Then

$$(4.2) \quad L^2u = L^2(u_+ + u_- + n) = L(\lambda u_+ - \lambda u_-) = \lambda^2(u_+ + u_-) = \mu u = \mu(u_+ + u_- + n).$$

From (4.2), we have

$$(4.3) \qquad (\lambda^2 - \mu)(u_+ + u_-) = \mu n;$$

so when $\mu \neq 0$, we have $n = 0$ and $\mu = \lambda^2$. We see that every nonzero eigenvalue of $L^2|_V$ is the square of an eigenvalue of L. (In fact, the same would be true even in the nonsemisimple case.)

Thus we can find the (nonzero) eigenvalues of L by finding the (nonzero) eigenvalues of $L^2|_V$ and taking their square roots. We define

$$(4.4) \qquad \mathcal{L} \equiv L^2|_V$$

and note that the eigenvalues of $\mathscr{L}$ must be real (since the eigenvalues of L are real or pure imaginary).

5. Eigenvalues of $\mathscr{L}$. Since $\gamma L^2 = L\gamma L = L^2\gamma$ for all $\gamma \in \Gamma$, we have

$$\hat{\mathscr{C}}\mathscr{L} = \mathscr{L}\hat{\mathscr{C}} \quad \text{and} \quad \hat{\mathscr{T}}\mathscr{L} = \mathscr{L}\hat{\mathscr{T}}. \tag{5.1}$$

This means that (in the basis $(z_1, z_2, \dots, z_n)$) advancing the rows of $\mathscr{L}$ by 1 is equivalent to advancing the columns by -1, and that exchanging the jth and $(n-j)$th rows of $\mathscr{L}$ for all j is equivalent to exchanging the jth and $(n-j)$th columns for all j. Hence

$$\mathscr{L} = \begin{pmatrix} L_n & L_1 & L_2 & \cdots & L_{n-1} \\ L_{n-1} & L_n & L_1 & \cdots & L_{n-2} \\ L_{n-2} & L_{n-1} & L_n & \cdots & L_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ L_1 & L_2 & L_3 & \cdots & L_n \end{pmatrix} \quad \text{with } L_j = L_{n-j}. \tag{5.2}$$

Direct calculation in this basis gives

$$L_j = \sum_{k=1}^{n} \left[\frac{\partial f_n^z}{\partial x_k}\bigg|_{r^0} \cdot \frac{\partial f_k^x}{\partial z_j}\bigg|_{r^0} + \frac{\partial f_n^z}{\partial y_k}\bigg|_{r^0} \cdot \frac{\partial f_k^y}{\partial z_j}\bigg|_{r^0} \right]. \tag{5.3}$$

We see by (5.2) that $\mathscr{L}$ is circulant. Hence $\mathscr{L}$ can be diagonalized by

$$U = \frac{1}{\sqrt{n}} \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{n-1} & \omega^{2(n-1)} & \cdots & \omega^{(n-1)^2} \end{pmatrix}, \qquad \omega = e^{i2\pi/n} \tag{5.4}$$

as

$$U^*\mathscr{L}U = \operatorname{diag}(p(1), p(\omega), \dots, p(\omega^{n-1})), \tag{5.5a}$$

where

$$p(\mu) = L_n + L_1\mu + \cdots + L_{n-1}\mu^{n-1} \tag{5.5b}$$

and U^* is the conjugate transpose of U (and equal to U^{-1}). (For diagonalization of circulant matrices, cf. [O].) Hence the eigenvalues of $\mathscr{L}$ are

$$p(1), p(\omega), \dots, p(\omega^{n-1}) \tag{5.6}$$

with p given by (5.5b).

6. Normal form for $F = (F^x, F^y, F^z)$. The calculations so far have depended only on the time reversibility and $D_n \times O(2)$ symmetry of the vector field $f(r)$ and thus apply to general problems with those symmetries. (We may replace the two-body interaction form (0.2) for f with the assumption that f commutes with an action of D_n fixing some horizontal equilibrium and then calculate eigenvalues exactly as described.) Here we obtain the general form of vector fields in our class up to linearization about r^0 using the full information we have on $f(r)$. We show that the entire class of problems may be divided into two families, parametrized by the degree of singularity β associated with the homogeneity of $f(r)$ and by the relative magnitudes of the vertical versus horizontal components of the vector field.

First, f is time-reversible (H1) if and only if

$$\begin{aligned} F^x(x, y, -z) &= -F^x(x, y, z), \\ F^y(x, y, -z) &= -F^y(x, y, z), \\ F^z(x, y, -z) &= F^z(x, y, z). \end{aligned} \tag{6.1}$$

Since we are concerned with the linearization of f about r^0, we may take as representative,

$$\begin{aligned} F^x(x, y, z) &= a(x, y)z, \\ F^y(x, y, z) &= b(x, y)z, \\ F^z(x, y, z) &= c(x, y). \end{aligned} \tag{6.2}$$

We must also incorporate the implications of $O(2)$ symmetry (H2) (cf. [GSS]). $SO(2)$ symmetry (commutation with rotations) implies (F^x, F^y) has length dependent oniy on z and $|(x, y)| = \sqrt{x^2+y^2}$, and direction given by a rotation ϕ of $\arg((x, y))$, where ϕ depends only on z. Moreover, F^z must depend only on z and $|(x, y)|$. Hence we have representative forms

$$\begin{aligned} a(x, y) &= g(x^2+y^2)(x \cos \phi - y \sin \phi), \\ b(x, y) &= g(x^2+y^2)(x \sin \phi + y \cos \phi), \\ c(x, y) &= c(x^2+y^2), \end{aligned} \tag{6.3}$$

where ϕ is a constant. In addition, we require flip symmetry in the (x, y)-plane. We should, therefore, take $\sin \phi = 0$. Finally, homogeneity (H3) forces the functions g and c to take particular monomial (possibly fractional powered) forms depending on β.

We have

$$f_j(r) = \sum_{k \neq j} \frac{F(r_j - r_k)}{|r_j - r_k|^\alpha} \tag{0.2}$$

with $F = (F^x, F^y, F^z)$ given by

$$\begin{aligned} F^x(x, y, z) &= g \cdot (x^2+y^2)^{\alpha/2-\beta/2-1} \cdot xz, \\ F^y(x, y, z) &= g \cdot (x^2+y^2)^{\alpha/2-\beta/2-1} \cdot yz, \\ F^z(x, y, z) &= c \cdot (x^2+y^2)^{\alpha/2-\beta/2}, \end{aligned} \tag{6.4}$$

where g and c are now constants. Although α appears in this form for $f(r)$, it is not a parameter since in linearization about r^0 the effects of α factor out of the numerator and denominator. Thus we arrive at our normal form for f:

$$f_j(r) = \sum_{k \neq j} \frac{F(r_j - r_k)}{|r_j - r_k|^{\beta+2}} \tag{0.2'}$$

with $F = (F^x, F^y, F^z)$ given by

$$\begin{aligned} F^x(x, y, z) &= g \cdot xz, \\ F^y(x, y, z) &= g \cdot yz, \\ F^z(x, y, z) &= c \cdot (x^2+y^2). \end{aligned} \tag{6.5}$$

We see that up to linearization, our class of vector field problems is parametrized by g/c and β. The class may be divided into two families according to the sign of g/c.

When $g/c > 0$, we say that the vertical and horizontal components of the vector field act "in concert"; when $g/c < 0$, we say they act "in opposition." The Stokeslet problem corresponds to $g/c = \beta = 1$, and the dipole example to $g/c = -3$, $\beta = 3$.

7. Eigenvalue calculation. We are now in a position to obtain formulas for the nonzero eigenvalues of the linearization of $f(r)$ at r^0 as functions of the singularity β and the constants g and c. Combining (5.3), (5.5b), and (6.5), and after much manipulation (see Appendix), we arrive at

$$\begin{aligned}\lambda_l^2 = -gc\beta \sum_{j\neq n}\sum_{k\neq n} &\left[\left(1+\cos\frac{2\pi kl}{n}\right)\left(1-\cos\frac{2\pi k}{n}\right)\left(1-\cos\frac{2\pi jl}{n}\right)\right.\\ &\left.\cdot\left(2\left(1-\cos\frac{2\pi k}{n}\right)\right)^{\beta/2+1}\left(2\left(1-\cos\frac{2\pi j}{n}\right)\right)^{\beta/2+1}\right]^{-1} \qquad (7.1)\\ &\cdot\left[\left(2\left(1-\cos\frac{2\pi k}{n}\right)\right)^{\beta/2+1}\left(2\left(1-\cos\frac{2\pi j}{n}\right)\right)^{\beta/2+1}\right]^{-1}\end{aligned}$$

or

$$\begin{aligned}\lambda_l^2 = -\frac{gc\beta}{4} \sum_{j\neq n}\sum_{k\neq n} &\left[\left(1+\cos\frac{2\pi kl}{n}\right)\left(1-\cos\frac{2\pi jl}{n}\right)-\left(1+\cos\frac{2\pi k}{n}\right)\right.\\ &\left.\cdot\, p_l\left(\cos\frac{2\pi k}{n}\right)\left(1+\cos\frac{2\pi j}{n}\right)p_l\left(\cos\frac{2\pi j}{n}\right)\right] \qquad (7.1')\\ &\cdot\left[\left(2\left(1-\cos\frac{2\pi k}{n}\right)\right)^{\beta/2}\left(2\left(1-\cos\frac{2\pi j}{n}\right)\right)^{\beta/2}\right]^{-1},\end{aligned}$$

where p_l is a notational device referring to the polynomial given by

$$\sin(lt) = \sin t \cdot p_l(\cos t). \qquad (7.2)$$

(It can be shown easily by induction on l that $\sin(lt)$ always has the form (7.2).) Hence up to sign and the scaling factor $|gc|$, the linearizations about r^0 in our class depend only on β. Throughout the rest of our discussion, we will take $gc = 1$. The implications for other values of gc are obvious.

Noting that $\lambda_l^2 = \lambda_{n-l}^2$ and $\lambda_0^2 = 0$, we see that it is enough to compute λ_l^2 for $l = 1, \ldots, \operatorname{int}(n/2)$. The double nature of these eigenvalues is as expected from our eigenvalue structure analysis. Figure 1 shows the dependence of λ_l^2 on β for a typical n.

8. Stability thresholds. Recall that by stability of a case in our vector field class we mean that the linearization of $f(r)$ at r^0 has only zero and pure imaginary eigenvalues, i.e., that λ_l^2 is nonpositive for all l. When λ_l^2 is positive for some l, then r^0 is clearly unstable under $f(r)$. We investigate the cases $\beta < 0$ and $\beta > 0$ separately.

We begin with the case $\beta < 0$. From (7.1') we have for $l = 1$ (with $gc = 1$),

$$\begin{aligned}\lambda_1^2 &= -\frac{\beta}{4}\sum_{j\neq n}\sum_{k\neq n}\left[\left(1+\cos\frac{2\pi k}{n}\right)\left(1-\cos\frac{2\pi j}{n}\right)-\left(1+\cos\frac{2\pi k}{n}\right)\left(1+\cos\frac{2\pi j}{n}\right)\right]\\ &\qquad\cdot\left[\left(2\left(1-\cos\frac{2\pi k}{n}\right)\right)^{\beta/2}\left(2\left(1-\cos\frac{2\pi j}{n}\right)\right)^{\beta/2}\right]^{-1} \qquad (8.1)\\ &= \frac{\beta}{2}\sum_{k\neq n}\left(1+\cos\frac{2\pi k}{n}\right)\Big/\left(2\left(1-\cos\frac{2\pi k}{n}\right)\right)^{\beta/2}\sum_{j\neq n}\cos\frac{2\pi j}{n}\Big/\left(2\left(1-\cos\frac{2\pi j}{n}\right)\right)^{\beta/2}.\end{aligned}$$

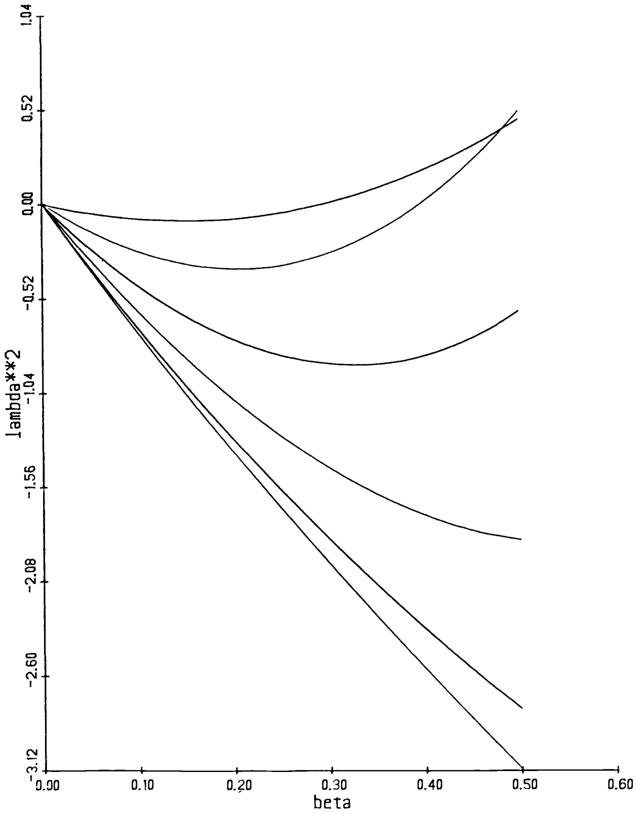


FIG. 1. λ_l^2 *as* β *varies. The case* $n=12$ *is shown, with curves for* $l=1,\ldots,6$ *appearing from top to bottom* $(g=c=1)$.

Since the sum over k in (8.1) is always positive and β is negative, the sign of λ_1^2 depends only on the sign of the sum over j in (8.1):

$$s(\beta, n)=\sum_{j\neq n}\frac{\cos(2\pi j/n)}{(2(1-\cos(2\pi j/n)))^{\beta/2}}. \tag{8.2}$$

Keeping n fixed, we first observe that as $\beta\to 0^-$, $s(\beta, n)\to -1$. Next, we note that all terms of $s(\beta, n)$ are nondecreasing functions of β except when j is such that $\pi/3<2\pi j/n<\pi/2$ or $2\pi-\pi/2<2\pi j/n<2\pi-\pi/3$. On the other hand, when $2\pi j/n$ falls in one of these regions, we can consider the sum of the jth and $2j$th (modulo n) terms. If we define

$$s_j(\beta, n)\equiv\frac{\cos(2\pi j/n)}{(2(1-\cos(2\pi j/n)))^{\beta/2}}+\frac{\cos(2\pi 2j/n)}{(2(1-\cos(2\pi 2j/n)))^{\beta/2}}, \tag{8.3}$$

then differentiating with respect to β and rearranging, we obtain

$$s_j'(\beta, n)>0 \tag{8.4}$$

if and only if

$$\beta<2\cdot\frac{\ln\left(\dfrac{\cos(4\pi j/n)\ln(2(1-\cos(4\pi j/n)))}{-\cos(2\pi j/n)\ln(2(1-\cos(2\pi j/n)))}\right)}{\ln(2(1+\cos(2\pi j/n)))}. \tag{8.5}$$

But the right-hand side of (8.5) is always positive so when $\beta<0$, (8.5) and hence (8.4) holds. We see that $s(\beta,n)$ is nondecreasing and approaches -1 as $\beta\to0^-$, so $s(\beta,n)$ is negative. Then we have $\lambda_1^2>0$ for all $\beta<0$ (any n) so that the $\beta<0$ case is always unstable. (Note that by "any n" we mean $n\geqq3$ since the cases $n=1,2$ are static in relative coordinates.)

Next we consider the $\beta>0$ case. We will require the identities

$$\begin{gathered}\sum_{k\neq n}1=n-1,\\ \sum_{k\neq n}\cos\frac{2\pi k}{n}=-1,\\ \sum_{k\neq n}\left(1+\cos\frac{2\pi k}{n}\right)p_l\left(\cos\frac{2\pi k}{n}\right)=n-2l.\end{gathered} \tag{8.6}$$

The first is obvious, the second is clear when we consider that $\cos(2\pi k/n)=\operatorname{Re}(\exp i(2\pi k/n))$, and the third is easily proved by induction on n. From the formulas for λ_l^2 (7.1′) and the identities (8.6), we have as $\beta\to0^+$,

$$\lambda_l^2\to-\frac{\beta}{4}[(n-1-1)(n-1+1)-(n-2l)^2]=\beta\left[n\left(\frac{1}{2}-l\right)+l^2\right], \tag{8.7}$$

so that $\lambda_l^2\to0^-$ as $\beta\to0^+$. On the other hand, as $\beta\to+\infty$, the dominant terms in the sum for λ_l^2 are those corresponding to $j,k=1,n-1$ so that

$$\lambda_l^2\to4\beta\frac{(1-\cos^2(2\pi l/n))\cos(2\pi/n)}{(2(1-\cos(2\pi/n)))^{\beta+1}} \tag{8.8}$$

as long as the right-hand side of (8.8) is nonzero (otherwise the next term must be examined). We see that as $\beta\to+\infty$, λ_l^2 is positive as long as $n>4$ (and $l\neq n/2$ for n even). Hence it is clear that for $n>4$, the stability threshold $\beta^*(n)$ exists. Studying the λ_l^2 numerically for $n\leqq20$, the first eigenvalue squared in each case to become positive as β increases is λ_1^2. It can be shown that λ_1^2 is zero for exactly one value of positive β by studying the behavior of $s(\beta,n)$ defined in (8.2). Table 1 shows approximate values of $\beta^*(n)$ for $4<n\leqq20$. For $n=5$, we have, in fact, $\beta^*(5)=2$ exactly:

TABLE 1
Stability thresholds.

n	$\beta^*(n)$	n	$\beta^*(n)$
5	2.000	13	0.258
6	1.121	14	0.232
7	0.772	15	0.210
8	0.585	16	0.192
9	0.470	17	0.176
10	0.391	18	0.163
11	0.334	19	0.152
12	0.292	20	0.142

writing $s(2,5)$ in terms of the golden number $\tau = 2\cos(\pi/5) = (1+\sqrt{5})/2$ (cf. [C]), it is easily seen that $s(2,5)=0$.

The case for $n \leqq 4$ is somewhat different. The property which distinguishes the cases $n \leqq 4$ from the cases $n > 4$ is that in the former, none of the $2\pi k/n$ have positive cosine while all in the latter have at least $\cos(2\pi/n) > 0$. In fact, the cases $n \leqq 4$ are stable for all $\beta > 0$. To see this, we simply calculate the eigenvalues explicitly for $n = 3, 4$ (the cases $n = 1, 2$ are static in relative coordinates). For $n = 3$, $\lambda_1^2 = -\beta/(2 \cdot 3^\beta)$ and for $n = 4$, $\lambda_1^2 = -\beta/2^{3\beta/2}$, $\lambda_2^2 = -2\beta/2^{3\beta/2}$. Hence we have stability for all $\beta > 0$.

In summary, we have the following.

THEOREM II (stability threshold). *Consider $dr/dt = f(r)$, where $f(r)$ has the form* (0.2) *and satisfies the symmetry relations* (H1)-(H3). *Then for $gc > 0$ (with g and c as appearing in* (6.5)):

(i) *For all $n \geqq 3$, r^0 is unstable when $\beta < 0$;*

(iia) *For $n = 3, 4$, r^0 is stable when $\beta > 0$;*

(iib) *For $n > 4$, there exists $\beta^*(n) < \infty$ such that r^0 is stable for all $0 < \beta < \beta^*(n)$ (take $\beta^*(n)$ to be the largest such β for which this is true).*

9. Nonresonant periodic solutions. We begin by stating the time-reversible, spatially symmetric version of the nonresonant Lyapunov center theorem essentially as it appears in [GKL]. Consider a system of ODEs

$$\frac{dx}{dt} = f(x), \qquad x \in \mathbf{R}^n \tag{9.1}$$

with an equilibrium at x_0. Assume there exists $R : \mathbf{R}^n \to \mathbf{R}^n$ such that

$$f(Rx) = -Rf(x), \qquad Rx_0 = x_0 \quad \text{(time reversibility);}$$

and a compact Lie group Γ which acts on $\mathbf{R}^n$ and such that

$$f(\gamma x) = \gamma f(x), \quad \gamma x_0 = x_0 \quad \forall \gamma \in \Gamma \quad \text{(spatial symmetry).}$$

Next, consider Σ a subgroup of $\Gamma \times S^1$, where $\Gamma \times S^1$ acts on the Banach space of 2π-periodic mappings $\mathbf{R} \to \mathbf{R}^n$ as

$$(\gamma, \theta) \cdot x(t) = \gamma x(t+\theta) \quad \forall (\gamma, \theta) \in \Gamma \times S^1.$$

A periodic solution $x(t)$ of (9.1) is said to have symmetry Σ when $(\gamma, \theta) \cdot x(t) = x(t)$ for all $(\gamma, \theta) \in \Sigma$. From [GKL] we have the following.

THEOREM. *Assume that the Γ-equivariant system* (9.1) *has a Γ-invariant equilibrium x_0. Assume*

$$\pm\omega_0 i \quad \textit{are nonzero eigenvalues of } (df)_{x_0}, \quad \textit{and}$$

$$k\omega_0 i \quad \textit{is not an eigenvalue of } (df)_{x_0} \quad \textit{for } k = 2, 3, \ldots.$$

Assume that the generalized eigenspace V_i corresponding to the eigenvalues $\pm\omega_0 i$ has the form

$$V_i = W \oplus W,$$

where Γ acts absolutely irreducibly on W. Assume finally that (9.1) *has time-reversal symmetry R fixing x_0 and that the subgroup $\Sigma \subseteq \Gamma \times S^1$ satisfies*

$$\dim \operatorname{Fix}(\Sigma) \cap V_i = 2,$$

$$R(\operatorname{Fix}(\Sigma) \cap V_i) = \operatorname{Fix}(\Sigma) \cap V_i,$$

$$R|\operatorname{Fix}(\Sigma) \cap \operatorname{Ker}(df)_{x_0} = I.$$

Then there exists an $m+1$-parameter family of periodic solutions to (9.1), *with period near $2\pi/\omega_0$ and symmetry Σ, where*

$$m=\dim \operatorname{Fix}(\Sigma)\cap \operatorname{Ker}(df)_{x_0}.$$

To apply this theorem to our class of problems, we need to check several conditions. First, for $0<\beta<\beta^*(n)$, we know all eigenvalues are pure imaginary, so we may take $\omega_0=|i\lambda_l|$, with any l from $1,\dots,\operatorname{int}(n/2)$. The nonresonance condition ($k\omega_0 i$ not an eigenvalue for $k=2,3,\dots$) must be checked but seems to hold generically in this class. That $V_i=W\oplus W$ holds can be seen by taking the isotypic decomposition (2.13) and arguing (by uniqueness of isotypic decompositions) that since the null part of any component W_k^3 is Γ-isomorphic to W_k, V_i must be of the required form. (For $l=n/2$, the relevant component is W_{--}^2, which is also of the required form.) We note finally that, apart from the null direction corresponding to vertical translation which may clearly be ignored, all zero eigenvalues of $(df)_{r^0}$ correspond to *horizontal* null directions so that $R|\operatorname{Fix}(\Sigma)\cap\operatorname{Ker}(df)_{r^0}=I$ holds for any Σ we choose. Hence to apply the theorem, it is sufficient to check nonresonance and to look for two-dimensional fixed point subspaces which are R-invariant. We obtain the following.

THEOREM III (nonresonant periodic solutions). *Consider $dr/dt=f(r)$, where $f(r)$ has the form* (0.2) *and satisfies the symmetry relations* (H1)-(H3). *Let $gc>0$ (with g and c as appearing in* (6.5)) *and let $0<\beta<\beta^*(n)$. Let λ be any nonzero eigenvalue of $(df)_{r^0}$ and $E_{\pm\lambda}$ the eigenspace corresponding to $\pm\lambda$. Assume the following nonresonance condition holds*:

$$k\lambda \quad \textit{is not an eigenvalue of } (df)_{r^0} \quad \textit{for } k=2,3,\dots.$$

Let Σ be any subgroup of $\Gamma\times S^1\cong D_n\times S^1$ such that

(i) $\dim\operatorname{Fix}(\Sigma)\cap E_{\pm\lambda}=2$;

(ii) $\operatorname{Fix}(\Sigma)\cap E_{\pm\lambda}$ *is left invariant by R*;

(iii) $\dim\operatorname{Fix}(\Sigma)\cap\operatorname{Ker}(df)_{r^0}=m.$

Then there exists an $m+1$-parameter family of periodic solutions to $dr/dt=f(r)$ with period near $2\pi/|i\lambda|$ and symmetry Σ.

The resulting symmetric families of periodic solutions are listed and described in [GKL] for $3\leqq n\leqq 6$ (there for the Stokeslet problem, but the results apply equally here). The relevant subgroups of $D_n\times S^1$ are listed in full in [GSS]. Independent calculations for the period of small orbits in the Stokeslet problem when $n=4$ for synchronous rhombi (corresponding here to the eigenvalues $\pm\lambda_2$) are performed in [TK].

10. Remarks. The focus here has been on exploiting symmetry properties to extract information about our class of problems. Several related vector field classes (e.g., inhomogeneous problems satisfying the other symmetry relations of our class, or $Z_n\times SO(2)$ rather than $D_n\times O(2)$) can be studied in essentially the same way. The information obtained here about the linearized problem provides the basis for nonlinear investigations such as the determination of families of near-equilibrium periodic solutions performed above.

Since physical problems tend to have integer values of the singularity strength β, and since $\beta^*(n)<1$ for $n>6$, we do not expect to find clusters of more than 6 bodies persisting in nature for problems of the type considered when vertical and horizontal components act in concert (e.g., in the sedimentation problem). In fact, when $\beta\neq 1$, under decay-type interaction, we expect only to find clusters of 3, 4, or 5 bodies. When the components act in opposition (as in the dipole example), we do not expect to find clusters of 3 or 4 bodies under decay-type interaction.

The parameter β may also be viewed as a bifurcation parameter for the family of decay-type problems in the class considered above. It can be shown that the family undergoes a time-reversible, equivariant pitchfork bifurcation as β passes through the critical value $\beta^*(n)$ (McComb, [M]).

Appendix: derivation of eigenvalue formulas. (Note: indices should be taken modulo n where appropriate.) The nonzero eigenvalues of the linearization of $f(r)$ at r^0 are given by (5.6):

$$\lambda_l^2 = p(\omega^l), \tag{A.0}$$

where $\omega = \exp i2\pi/n$ and

$$p(\mu) = L_n + L_1\mu + \cdots + L_{n-1}\mu^{n-1} \tag{5.5b}$$

with

$$L_j = \sum_{k=1}^{n} \left[\frac{\partial f_n^z}{\partial x_k}\bigg|_{r^0} \cdot \frac{\partial f_k^x}{\partial z_j}\bigg|_{r^0} + \frac{\partial f_n^z}{\partial y_k}\bigg|_{r^0} \cdot \frac{\partial f_k^y}{\partial z_j}\bigg|_{r^0} \right]. \tag{5.3}$$

If we define

$$A_{13}(i,j) = \frac{\partial f_i^x}{\partial z_j}\bigg|_{r^0}, \quad A_{23}(i,j) = \frac{\partial f_i^y}{\partial z_j}\bigg|_{r^0}, \quad A_{31}(i,j) = \frac{\partial f_i^z}{\partial x_j}\bigg|_{r^0}, \quad A_{32}(i,j) = \frac{\partial f_i^z}{\partial y_j}\bigg|_{r^0}, \tag{A.1}$$

we have

$$L_j = \sum_{k=1}^{n} [A_{31}(n,k)A_{13}(k,j) + A_{32}(n,k)A_{23}(k,j)]. \tag{5.3'}$$

Taking the appropriate derivatives and substituting in r^0, we have, for $k \neq n$,

$$\begin{aligned} A_{31}(n,k) &= \frac{\beta c(1-\cos(2\pi k/n))}{(2(1-\cos(2\pi k/n)))^{\beta/2+1}}, \\ A_{31}(n,n) &= -\sum_{k\neq n} A_{31}(n,k), \\ A_{32}(n,k) &= \frac{\beta c(-\sin(2\pi k/n))}{(2(1-\cos(2\pi k/n)))^{\beta/2+1}}, \\ A_{32}(n,n) &= -\sum_{k\neq n} A_{32}(n,k), \end{aligned} \tag{A.2}$$

and for $k \neq j$,

$$\begin{aligned} A_{13}(k,j) &= \frac{g(\cos(2\pi j/n) - \cos(2\pi k/n))}{(2(1-\cos(2\pi(j-k)/n)))^{\beta/2+1}}, \\ A_{13}(j,j) &= -\sum_{k\neq j} A_{13}(j,k) = \sum_{k\neq j} A_{13}(k,j), \\ A_{23}(k,j) &= \frac{g(\sin(2\pi j/n) - \sin(2\pi k/n))}{(2(1-\cos(2\pi(j-k)/n)))^{\beta/2+1}}, \\ A_{23}(j,j) &= -\sum_{k\neq j} A_{23}(j,k) = \sum_{k\neq j} A_{23}(k,j). \end{aligned} \tag{A.3}$$

We see that for $j \neq n$, we have

$$
\begin{aligned}
L_j &= \sum_{k=1}^{n} [A_{31}(n,k)A_{13}(k,j) + A_{32}(n,k)A_{23}(k,j)] \\
&= \sum_{k \neq j,n} [A_{31}(n,k)A_{13}(k,j) + A_{32}(n,k)A_{23}(k,j)] \\
&\quad + \sum_{k \neq j} [A_{31}(n,j)A_{13}(k,j) + A_{32}(n,j)A_{23}(k,j)] \\
&\quad - \sum_{k \neq n} [A_{31}(n,k)A_{13}(n,j) + A_{32}(n,k)A_{23}(n,j)] \\
&= \sum_{k \neq j,n} [A_{31}(n,k)A_{13}(k,j) + A_{32}(n,k)A_{23}(k,j)] \\
&\quad + \sum_{k \neq j,n} [A_{31}(n,j)A_{13}(k,j) + A_{32}(n,j)A_{23}(k,j)] \\
&\quad + [A_{31}(n,j)A_{13}(n,j) + A_{32}(n,j)A_{23}(n,j)] \\
&\quad - \sum_{k \neq j,n} [A_{31}(n,k)A_{13}(n,j) + A_{32}(n,k)A_{23}(n,j)] \\
&\quad - [A_{31}(n,j)A_{13}(n,j) + A_{32}(n,j)A_{23}(n,j)] \\
&= \sum_{k \neq j,n} [A_{31}(n,k)A_{13}(k,j) + A_{32}(n,k)A_{23}(k,j)] \\
&\quad + \sum_{k \neq j,n} [A_{31}(n,j)A_{13}(k,j) + A_{32}(n,j)A_{23}(k,j)] \\
&\quad - \sum_{k \neq j,n} [A_{31}(n,k)A_{13}(n,j) + A_{32}(n,k)A_{23}(n,j)].
\end{aligned}
\tag{A.4}
$$

We also have

$$
\begin{aligned}
L_n &= \sum_{k=1}^{n} [A_{31}(n,k)A_{13}(k,n) + A_{32}(n,k)A_{23}(k,n)] \\
&= \sum_{k \neq n} [A_{31}(n,k)A_{13}(k,n) + A_{32}(n,k)A_{23}(k,n)] \\
&\quad + \sum_{k \neq n} A_{31}(n,k) \sum_{j \neq n} A_{13}(n,j) + \sum_{k \neq n} A_{32}(n,k) \sum_{j \neq n} A_{23}(n,j) \\
&= \sum_{k \neq n} \Bigg[A_{31}(n,k)\Big(A_{13}(k,n) + \sum_{j \neq n} A_{13}(n,j) \Big) \\
&\qquad + A_{32}(n,k)\Big(A_{23}(k,n) + \sum_{j \neq n} A_{23}(n,j) \Big) \Bigg] \\
&= \sum_{k \neq n} \sum_{j \neq k,n} [A_{31}(n,k)A_{13}(n,j) + A_{32}(n,k)A_{23}(n,j)] \\
&= \sum_{j \neq n} \sum_{k \neq j,n} [A_{31}(n,k)A_{13}(n,j) + A_{32}(n,k)A_{23}(n,j)].
\end{aligned}
\tag{A.5}
$$

Noting that

$$
\begin{aligned}
\sum_{j \neq n} L_j &= \sum_{j \neq n} \sum_{k \neq j,n} [A_{31}(n,k)A_{13}(k,j) + A_{32}(n,k)A_{23}(k,j)] \\
&\quad + \sum_{j \neq n} \sum_{k \neq j,n} [A_{31}(n,j)A_{13}(k,j) + A_{32}(n,j)A_{23}(k,j)] - L_n \\
&= \sum_{j \neq n} \sum_{k \neq j,n} [A_{31}(n,k)A_{13}(k,j) + A_{32}(n,k)A_{23}(k,j)] \\
&\quad - \sum_{k \neq n} \sum_{j \neq k,n} [A_{31}(n,j)A_{13}(j,k) + A_{32}(n,j)A_{23}(j,k)] - L_n \\
&= -L_n,
\end{aligned}
\tag{A.6}
$$

we see that the eigenvalues we seek are

$$\lambda_l^2 \equiv p(\omega^l) = -\sum_{j=1}^{n} L_j\left(1-\cos\frac{2\pi jl}{n}\right) + i\sum_{j=1}^{n} L_j\left(\sin\frac{2\pi jl}{n}\right). \tag{A.7}$$

But $L_j = L_{n-j}$ so $\operatorname{Im}(\lambda_l^2) = 0$ (as expected). Hence we have

$$\lambda_l^2 = -\sum_{j=1}^{n} L_j\left(1-\cos\frac{2\pi jl}{n}\right), \tag{A.8}$$

where L_j is given by (A.4). Note that $\lambda_l^2 = \lambda_{n-l}^2$ and $\lambda_0^2 = 0$.

Substituting (A.4) into (A.8) we obtain

$$\begin{aligned}
\lambda_l^2 &= -\sum_{j\neq n} L_j\left(1-\cos\frac{2\pi jl}{n}\right)\\
&= -\sum_{j\neq n}\sum_{k\neq j,n}[A_{31}(n,k)A_{13}(k,j)+A_{32}(n,k)A_{23}(k,j)]\left(1-\cos\frac{2\pi jl}{n}\right)\\
&\quad -\sum_{j\neq n}\sum_{k\neq j,n}[A_{31}(n,j)A_{13}(k,j)+A_{32}(n,j)A_{23}(k,j)]\left(1-\cos\frac{2\pi jl}{n}\right)\\
&\quad +\sum_{j\neq n}\sum_{k\neq j,n}[A_{31}(n,k)A_{13}(n,j)+A_{32}(n,k)A_{23}(n,j)]\left(1-\cos\frac{2\pi jl}{n}\right)
\end{aligned}$$

$$\begin{aligned}
&= -\sum_{j\neq n}\sum_{k\neq j,n}[A_{31}(n,k)A_{13}(k,j)+A_{32}(n,k)A_{23}(k,j)]\left(1-\cos\frac{2\pi jl}{n}\right)\\
&\quad +\sum_{k\neq n}\sum_{j\neq k,n}[A_{31}(n,k)A_{13}(k,j)+A_{32}(n,k)A_{23}(k,j)]\left(1-\cos\frac{2\pi kl}{n}\right)\\
&\quad +\sum_{j\neq n}\sum_{k\neq j,n}[A_{31}(n,k)A_{13}(n,j)+A_{32}(n,k)A_{23}(n,j)]\left(1-\cos\frac{2\pi jl}{n}\right)\\
&= \sum_{j\neq n}\sum_{k\neq j,n}[A_{31}(n,k)A_{13}(k,j)+A_{32}(n,k)A_{23}(k,j)]\left(\cos\frac{2\pi jl}{n}-\cos\frac{2\pi kl}{n}\right)\\
&\quad +\sum_{j\neq n}\sum_{k\neq j,n}[A_{31}(n,k)A_{13}(n,j)+A_{32}(n,k)A_{23}(n,j)]\left(1-\cos\frac{2\pi jl}{n}\right).
\end{aligned} \tag{A.9}$$

Now substituting according to (A.2)-(A.3),

$$\begin{aligned}
\lambda_l^2 = gc\beta\sum_{j\neq n}\sum_{k\neq j,n}&\Bigg\{\left[\left(1-\cos\frac{2\pi k}{n}\right)\left(\cos\frac{2\pi j}{n}-\cos\frac{2\pi k}{n}\right)-\sin\frac{2\pi k}{n}\left(\sin\frac{2\pi j}{n}-\sin\frac{2\pi k}{n}\right)\right]\\
&\cdot\left[\left(2\left(1-\cos\frac{2\pi k}{n}\right)\right)^{\beta/2+1}\left(2\left(1-\cos\frac{2\pi(j-k)}{n}\right)\right)^{\beta/2+1}\right]^{-1}\Bigg\}\\
&\cdot\left(\cos\frac{2\pi jl}{n}-\cos\frac{2\pi kl}{n}\right)\\
-gc\beta\sum_{j\neq n}\sum_{k\neq j,n}&\Bigg\{\left[\left(1-\cos\frac{2\pi k}{n}\right)\left(1-\cos\frac{2\pi j}{n}\right)+\sin\frac{2\pi k}{n}\sin\frac{2\pi j}{n}\right]\\
&\cdot\left[\left(2\left(1-\cos\frac{2\pi k}{n}\right)\right)^{\beta/2+1}\left(2\left(1-\cos\frac{2\pi(j-k)}{n}\right)\right)^{\beta/2+1}\right]^{-1}\Bigg\}\\
&\cdot\left(1-\cos\frac{2\pi jl}{n}\right).
\end{aligned} \tag{A.10}$$

Simplifying (A.10),

(A.11)

$$
\begin{aligned}
\lambda_l^2 = gc\beta \sum_{j\neq n}\ \sum_{k\neq j,n} & \left[\left(\cos\frac{2\pi j}{n}-\cos\frac{2\pi k}{n}\right)+\left(1-\cos\frac{2\pi(j-k)}{n}\right)\right]\left(\cos\frac{2\pi jl}{n}-\cos\frac{2\pi kl}{n}\right)\\
& \cdot\left[\left(2\left(1-\cos\frac{2\pi k}{n}\right)\right)^{\beta/2+1}\left(2\left(1-\cos\frac{2\pi(j-k)}{n}\right)\right)^{\beta/2+1}\right]^{-1}\\
-gc\beta \sum_{j\neq n}\ \sum_{k\neq j,n} & \left[\left(\cos\frac{2\pi(j-k)}{n}-\cos\frac{2\pi k}{n}\right)+\left(1-\cos\frac{2\pi j}{n}\right)\right]\left(1-\cos\frac{2\pi jl}{n}\right)\\
& \cdot\left[\left(2\left(1-\cos\frac{2\pi k}{n}\right)\right)^{\beta/2+1}\left(2\left(1-\cos\frac{2\pi j}{n}\right)\right)^{\beta/2+1}\right]^{-1}\\
= gc\beta \sum_{k\neq n}\ \sum_{\substack{k-j\\ \neq k,n}} & \left[\left(\cos\frac{2\pi[k-(k-j)]}{n}-\cos\frac{2\pi k}{n}\right)+\left(1-\cos\frac{2\pi(k-j)}{n}\right)\right]\\
& \cdot\left(\cos\frac{2\pi[k-(k-j)]l}{n}-\cos\frac{2\pi kl}{n}\right)\\
& \cdot\left[\left(2\left(1-\cos\frac{2\pi k}{n}\right)\right)^{\beta/2+1}\left(2\left(1-\cos\frac{2\pi(k-j)}{n}\right)\right)^{\beta/2+1}\right]^{-1}\\
-gc\beta \sum_{k\neq n}\ \sum_{j\neq k,n} & \left[\left(\cos\frac{2\pi(k-j)}{n}-\cos\frac{2\pi k}{n}\right)+\left(1-\cos\frac{2\pi j}{n}\right)\right]\left(1-\cos\frac{2\pi jl}{n}\right)\\
& \cdot\left[\left(2\left(1-\cos\frac{2\pi k}{n}\right)\right)^{\beta/2+1}\left(2\left(1-\cos\frac{2\pi j}{n}\right)\right)^{\beta/2+1}\right]^{-1}\\
= gc\beta \sum_{k\neq n}\ \sum_{j\neq k,n} & \left[\left(\cos\frac{2\pi(j-k)}{n}-\cos\frac{2\pi k}{n}\right)+\left(1-\cos\frac{2\pi j}{n}\right)\right]\\
& \cdot\left[\left(\cos\frac{2\pi(j-k)l}{n}-\cos\frac{2\pi kl}{n}\right)-\left(1-\cos\frac{2\pi jl}{n}\right)\right]\\
& \cdot\left[\left(2\left(1-\cos\frac{2\pi k}{n}\right)\right)^{\beta/2+1}\left(2\left(1-\cos\frac{2\pi j}{n}\right)\right)^{\beta/2+1}\right]^{-1}\\
= gc\beta \sum_{k\neq n}\ \sum_{j\neq n} & \left[\left(\cos\frac{2\pi(j-k)}{n}-\cos\frac{2\pi k}{n}\right)+\left(1-\cos\frac{2\pi j}{n}\right)\right]\\
& \cdot\left[\left(\cos\frac{2\pi(j-k)l}{n}-\cos\frac{2\pi kl}{n}\right)-\left(1-\cos\frac{2\pi jl}{n}\right)\right]\\
& \cdot\left[\left(2\left(1-\cos\frac{2\pi k}{n}\right)\right)^{\beta/2+1}\left(2\left(1-\cos\frac{2\pi j}{n}\right)\right)^{\beta/2+1}\right]^{-1}.
\end{aligned}
$$

Defining

$$\begin{aligned} C(k)&=\frac{\beta c}{(2(1-\cos(2\pi k/n)))^{\beta/2+1}},\\ G(j)&=\frac{g}{(2(1-\cos(2\pi j/n)))^{\beta/2+1}}, \end{aligned}\tag{A.12}$$

we have

$$\begin{aligned} \lambda_l^2&=\sum_{j\neq n}\sum_{k\neq n}C(k)G(j)\left[\left(\cos\frac{2\pi(j-k)l}{n}-\cos\frac{2\pi kl}{n}\right)-\left(1-\cos\frac{2\pi jl}{n}\right)\right]\\ &\quad\cdot\left[\left(\cos\frac{2\pi(j-k)}{n}-\cos\frac{2\pi k}{n}\right)+\left(1-\cos\frac{2\pi j}{n}\right)\right]\\ &=\sum_{j\neq n}\sum_{k\neq n}C(k)G(j)\\ &\quad\cdot\left[\cos\frac{2\pi kl}{n}\cos\frac{2\pi jl}{n}+\sin\frac{2\pi kl}{n}\sin\frac{2\pi jl}{n}-\cos\frac{2\pi kl}{n}-\left(1-\cos\frac{2\pi jl}{n}\right)\right]\\ &\quad\cdot\left[\cos\frac{2\pi k}{n}\cos\frac{2\pi j}{n}+\sin\frac{2\pi k}{n}\sin\frac{2\pi j}{n}-\cos\frac{2\pi k}{n}+\left(1-\cos\frac{2\pi j}{n}\right)\right]\\ &=\sum_{j\neq n}\sum_{k\neq n}C(k)G(j)\\ &\quad\cdot\left\{\cos\frac{2\pi kl}{n}\cos\frac{2\pi jl}{n}\cos\frac{2\pi k}{n}\cos\frac{2\pi j}{n}+\cos\frac{2\pi kl}{n}\cos\frac{2\pi jl}{n}\sin\frac{2\pi k}{n}\sin\frac{2\pi j}{n}\right.\\ &\qquad-\cos\frac{2\pi kl}{n}\cos\frac{2\pi jl}{n}\cos\frac{2\pi k}{n}+\cos\frac{2\pi kl}{n}\cos\frac{2\pi jl}{n}\left(1-\cos\frac{2\pi j}{n}\right)\\ &\qquad+\sin\frac{2\pi kl}{n}\sin\frac{2\pi jl}{n}\cos\frac{2\pi k}{n}\cos\frac{2\pi j}{n}+\sin\frac{2\pi kl}{n}\sin\frac{2\pi jl}{n}\sin\frac{2\pi k}{n}\sin\frac{2\pi j}{n}\\ &\qquad-\sin\frac{2\pi kl}{n}\sin\frac{2\pi jl}{n}\cos\frac{2\pi k}{n}+\sin\frac{2\pi kl}{n}\sin\frac{2\pi jl}{n}\left(1-\cos\frac{2\pi j}{n}\right)\\ &\qquad-\cos\frac{2\pi kl}{n}\cos\frac{2\pi k}{n}\cos\frac{2\pi j}{n}-\cos\frac{2\pi kl}{n}\sin\frac{2\pi k}{n}\sin\frac{2\pi j}{n}\\ &\qquad+\cos\frac{2\pi kl}{n}\cos\frac{2\pi k}{n}-\cos\frac{2\pi kl}{n}\left(1-\cos\frac{2\pi j}{n}\right)\\ &\qquad-\left(1-\cos\frac{2\pi jl}{n}\right)\cos\frac{2\pi k}{n}\cos\frac{2\pi j}{n}-\left(1-\cos\frac{2\pi jl}{n}\right)\sin\frac{2\pi j}{n}\sin\frac{2\pi k}{n}\\ &\qquad\left.+\left(1-\cos\frac{2\pi jl}{n}\right)\cos\frac{2\pi k}{n}-\left(1-\cos\frac{2\pi jl}{n}\right)\left(1-\cos\frac{2\pi j}{n}\right)\right\}. \end{aligned}\tag{A.13}$$

By eliminating terms which are odd in k (since they sum to zero over $k \neq n$) and rearranging,

(A.14)

$$\begin{aligned}\lambda_l^2 = \sum_{j\neq n}\sum_{k\neq n} C(k)G(j)\Bigg\{ &\cos\frac{2\pi kl}{n}\cos\frac{2\pi k}{n}\left(\cos\frac{2\pi jl}{n}\cos\frac{2\pi j}{n} - \cos\frac{2\pi jl}{n} - \cos\frac{2\pi j}{n} + 1\right) \\ &+ \cos\frac{2\pi kl}{n}\left(\cos\frac{2\pi jl}{n}\left(1-\cos\frac{2\pi j}{n}\right) - \left(1-\cos\frac{2\pi j}{n}\right)\right) \\ &+ \cos\frac{2\pi k}{n}\left(-\left(1-\cos\frac{2\pi jl}{n}\right)\cos\frac{2\pi j}{n} + \left(1-\cos\frac{2\pi jl}{n}\right)\right) \\ &-\left(1-\cos\frac{2\pi jl}{n}\right)\left(1-\cos\frac{2\pi j}{n}\right) \\ &+\sin\frac{2\pi kl}{n}\sin\frac{2\pi jl}{n}\sin\frac{2\pi k}{n}\sin\frac{2\pi j}{n}\Bigg\}, \\ = -\sum_{j\neq n}\sum_{k\neq n} C(k)G(j)\Bigg\{ &\left(1+\cos\frac{2\pi kl}{n} - \cos\frac{2\pi k}{n} - \cos\frac{2\pi kl}{n}\cos\frac{2\pi k}{n}\right) \\ &\cdot\left(1-\cos\frac{2\pi jl}{n}\right)\left(1-\cos\frac{2\pi j}{n}\right) \\ &-\sin\frac{2\pi kl}{n}\sin\frac{2\pi jl}{n}\sin\frac{2\pi k}{n}\sin\frac{2\pi j}{n}\Bigg\}, \\ = -\sum_{j\neq n}\sum_{k\neq n} C(k)G(j)\Bigg\{ &\left(1+\cos\frac{2\pi kl}{n}\right)\left(1-\cos\frac{2\pi k}{n}\right)\left(1-\cos\frac{2\pi jl}{n}\right)\left(1-\cos\frac{2\pi j}{n}\right) \\ &-\sin\frac{2\pi kl}{n}\sin\frac{2\pi k}{n}\sin\frac{2\pi jl}{n}\sin\frac{2\pi j}{n}\Bigg\}.\end{aligned}$$

Hence the nonzero eigenvalues of the linearization of f at r^0 are λ_l such that

(7.1)

$$\begin{aligned}\overset{2}{\lambda_l} = -gc\beta &\sum_{j\neq n}\sum_{k\neq n} \\ &\cdot\left[\left(1+\cos\left(\frac{2\pi kl}{n}\right)\right)\left(1-\cos\frac{2\pi k}{n}\right)\left(1-\cos\frac{2\pi jl}{n}\right)\right. \\ &\left.\cdot\left(1-\cos\frac{2\pi j}{n}\right) - \sin\frac{2\pi kl}{n}\sin\frac{2\pi k}{n}\sin\frac{2\pi jl}{n}\sin\frac{2\pi j}{n}\right] \\ &\cdot\left[\left(2\left(1-\cos\frac{2\pi k}{n}\right)\right)^{\beta/2+1}\left(2\left(1-\cos\frac{2\pi j}{n}\right)\right)^{\beta/2+1}\right]^{-1}\end{aligned}$$

REFERENCES

[C] H. S. M. COXETER, *Introduction to Geometry*, 2nd ed., John Wiley and Sons, Inc., New York, 1989.

[CLLS] R. E. CAFLISCH, C. LIM, J. H. C. LUKE, AND A. S. SANGANI, *Periodic solutions for three sedimenting spheres*, Phys. Fluids, 31 (1988), pp. 3175-3179.

[GKL] M. GOLUBITSKY, M. KRUPA, AND C. LIM, *Time-reversibility and particle sedimentation*, SIAM J. Appl. Math., 51 (1990), pp. 49-72.

[GSS] M. GOLUBITSKY, I. N. STEWART, AND D. G. SCHAEFFER, *Singularities and Groups in Bifurcation Theory*: *Vol.* II, Appl. Math. Sci., 69, Springer-Verlag, New York, 1988.

[H] L. M. HOCKING, *The behaviour of clusters of spheres falling in a viscous fluid. Part* 2. *Slow motion theory*, J. Fluid Mech., 20 (1964), pp. 129-139.

[JMS] K. O. L. F. JAYAWEERA, B. J. MASON, AND G. W. SLACK, *The behaviour of clusters of spheres falling in a viscous fluid. Part* 1. *Experiment*, J. Fluid Mech., 20 (1964), pp. 121-128.

[M] I. MCCOMB, *Generic pitchfork bifurcations of time-reversible, equivariant vector field families*, preprint.

[O] J. M. ORTEGA, *Matrix Theory*, University Series in Mathematics, Plenum Press, New York, 1987.

[S] M. SEVRYUK, *Reversible Systems*, Lecture Notes in Math., 1211, Springer-Verlag, Berlin, 1986.

[TK] E. M. TORY AND M. T. KAMEL, *A note on the periodic motion of four spheres*, Powder Technology, to appear.

SIAM J. MATH. ANAL.
Vol. 24, No. 4, pp. 1030–1042, July 1993

010

UNCONDITIONAL BASES OF WAVELETS FOR SOBOLEV SPACES*

GUSTAF GRIPENBERG[†]

Abstract. It is shown that an orthonormal wavelet basis for $L^2(\mathbb{R})$ associated with a multiresolution is an unconditional basis for the Sobolev space $\mathcal{H}^t(\mathbb{R})$, $t \in [-s, s]$, provided the father wavelet belongs to $\mathcal{H}^s(\mathbb{R})$ and the square of its absolute value has finite moments. A criterion for when the wavelet belongs to $\mathcal{H}^s(\mathbb{R})$ is given.

Key words. basis, unconditional, Sobolev space, wavelet, multiresolution

AMS subject classifications. primary 42C15; secondary 46B15, 46E35, 39B99

1. Introduction.

The purpose of this paper is to extend some of the results in [13] on unconditional bases for the Sobolev space $\mathcal{H}^s(\mathbb{R}; \mathbb{C})$, which are of the form $\{\psi(2^m \bullet -k)\}_{m,k\in\mathbb{Z}}$ where ψ is a mother wavelet, that is, $\{2^{m/2}\psi(2^m \bullet -k)\}_{m,k\in\mathbb{Z}}$ is an orthonormal basis for $L^2(\mathbb{R}; \mathbb{C})$. (Here $\bullet$ denotes a generic argument.) The analysis in this paper is restricted to the one-dimensional case, where there is also a father wavelet φ such that $V_{m+1} = V_m \oplus W_m$, where V_m is the space spanned by $\{\varphi(2^m \bullet -k)\}_{k\in\mathbb{Z}}$ and W_m is the space spanned by $\{\psi(2^m \bullet -k)\}_{k\in\mathbb{Z}}$. In other words, the wavelets are obtained from a multiresolution.

It is proved below that if ψ and φ belong to $\mathcal{H}^s(\mathbb{R}; \mathbb{C})$, where $s \geq 0$ and both functions decay sufficiently rapidly (e.g., all moments of the square of the absolute value are finite), then we get an unconditional basis for $\mathcal{H}^t(\mathbb{R}; \mathbb{C})$ for all $t \in [-s, s]$. In order to obtain this result one has to extend a known criterion (see [6], [15]) for when φ, $\psi \in \mathcal{H}^s(\mathbb{R}; \mathbb{C})$, to the case where the functions are not necessarily compactly supported. This criterion involves the spectral radius of an operator, and in the case of compactly supported wavelets the problem is reduced to calculating the spectral radius of a matrix. This means, for example, that one knows exactly for which positive values of s the wavelets constructed in [3] give rise to unconditional bases in $\mathcal{H}^s(\mathbb{R}; \mathbb{C})$.

The main point of this paper is not that it is possible to construct an unconditional wavelet basis, because this is well known and holds true, as can be seen from [13], for a large number of other types of functional spaces too, but to find out exactly what assumptions are needed on the wavelets. It turns out that the hypotheses one has to use are quite natural, and this constitutes additional evidence for the claim that wavelets are an extremely useful tool.

2. Statement of results.

First we have to define what we mean by a multiresolution or a multiresolution analysis as it is often called. We say that $(\{V_m\}_{m\in\mathbb{Z}}, \varphi)$ is a multiresolution of $L^2(\mathbb{R}; \mathbb{C})$ provided that the following four conditions hold.

$$\varphi \in L^2(\mathbb{R}; \mathbb{C}) \text{ and } V_m \text{ is, for each } m \in \mathbb{Z}, \text{ the closed subspace of } L^2(\mathbb{R}; \mathbb{C}) \text{ spanned by } \{\varphi(2^m \bullet -k)\}_{k\in\mathbb{Z}}, \tag{1}$$

*Received by the editors January 7, 1992; accepted for publication (in revised form) November 9, 1992.

[†]Department of Mathematics, University of Helsinki, Regeringsgatan 15, 00100 Helsingfors, Finland (Gustaf.Gripenberg@Helsinki.Fi).

$$V_m \subset V_{m+1}, \qquad m \in \mathbb{Z}, \tag{2}$$

(3) $\lim_{m\to\infty} P_m f = f$ for every $f \in L^2(\mathbb{R};\mathbb{C})$, where P_m is the orthogonal projection of $L^2(\mathbb{R};\mathbb{C})$ onto V_m,

(4) $\{\varphi(\bullet - k)\}_{k\in\mathbb{Z}}$ is an orthonormal set in $L^2(\mathbb{R};\mathbb{C})$.

The function φ is then said to be the father wavelet or scaling function.

The definition of a multiresolution is often given in a slightly different form (but with exactly the same content); see, e.g., [1], [4], and [11]–[13], so that the fact that $\{\varphi(2^m \bullet -k)\}_{k\in\mathbb{Z}}$ spans V_m is a consequence of the other assumptions. Condition (3) is often formulated as the requirement that $\bigcup_{m=-\infty}^{\infty} V_m$ is dense in $L^2(\mathbb{R};\mathbb{C})$ and it is combined with the assumption that $\bigcap_{m=-\infty}^{\infty} V_m = \{0\}$, which follows from the other conditions; see [1, p. 443] and note that the moment conditions assumed there are not used in the proof of this statement. It is not really essential that $\{\varphi(\bullet - k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis; it would suffice (as is often done) to require that it is an unconditional basis in which case it is a Riesz basis (see [8, Thm. 2.2.2]), but from such a basis one easily constructs an orthonormal one.

Since $\varphi \in V_0 \subset V_1$ it follows that φ can be expressed in terms of the functions $\{\varphi(2\bullet -k)\}_{k\in\mathbb{Z}}$ which span V_1, that is,

$$\varphi = 2\sum_{k\in\mathbb{Z}} \alpha(k)\varphi(2\bullet -k), \tag{5}$$

where

$$\alpha(k) = \int_{\mathbb{R}} \varphi(x)\overline{\varphi(2x-k)}\,\mathrm{d}x, \qquad k \in \mathbb{Z}.$$

We call this sequence α the filter associated with the multiresolution and it turns out to be crucial for the analysis and in particular for computations involving the wavelets, cf. [14].

If we take Fourier transforms of both sides of (5), then we get

$$\hat{\varphi}(2\bullet) = \hat{\alpha}(\bullet)\hat{\varphi}(\bullet). \tag{6}$$

From this equation one sees the advantages of having the normalizing factor 2 in (5). Here we have defined the Fourier transform of φ to be $\hat{\varphi} = \int_{\mathbb{R}} \mathrm{e}^{-2\pi\mathrm{i}\bullet x}\varphi(x)\,dx$ and the Fourier transform of α to be $\hat{\alpha} = \sum_{k\in\mathbb{Z}} \mathrm{e}^{-2\pi\mathrm{i}\bullet k}\alpha(k)$. Thus $\hat{\alpha}$ is periodic with period 1. It is possible to prove (see, e.g., [11, p. 31]) that

$$|\hat{\alpha}(\bullet)|^2 + \left|\hat{\alpha}\left(\bullet + \tfrac{1}{2}\right)\right|^2 = 1 \quad \text{a.e. on } \mathbb{R}. \tag{7}$$

Sequences α that satisfy (7) are called conjugate quadrature filters.

Having found the filter α we can define the mother wavelet ψ as follows:

$$\psi = 2\sum_{k\in\mathbb{Z}} (-1)^k \overline{\alpha(1-k)}\varphi(2\bullet -k).$$

For the Fourier transforms this implies that

$$\hat{\psi}(2\bullet) = -\mathrm{e}^{-2\pi\mathrm{i}\bullet}\overline{\hat{\alpha}\left(\bullet + \tfrac{1}{2}\right)}\hat{\varphi}(\bullet). \tag{8}$$

It follows from these definitions that the sets

$$\{2^{m/2}\psi(2^m \bullet -k)\}_{m,k\in\mathbb{Z}},$$

and

$$\{2^{m_0/2}\varphi(2^{m_0} \bullet -k), 2^{m/2}\psi(2^m \bullet -k)\}_{m\geq m_0,k\in\mathbb{Z}},$$

where $m_0 \in \mathbb{Z}$ is arbitrary, are orthonormal bases for $L^2(\mathbb{R};\mathbb{C})$; see, e.g., [4] or [13].

The Sobolev space $\mathcal{H}^s(\mathbb{R};\mathbb{C})$ consists of all functions f (or tempered distributions in the case where $s < 0$) such that

$$\|f\|^2_{\mathcal{H}^s(\mathbb{R})} = \int_{\mathbb{R}} (1+|\omega|^2)^s |\hat{f}(\omega)|^2 \,\mathrm{d}\omega < \infty.$$

If s is a nonnegative integer N, then $\mathcal{H}^N(\mathbb{R};\mathbb{C})$ consists of all square integrable functions f that are $N-1$ times continuously differentiable such that $f^{(N-1)}$ is locally absolutely continuous with a square integrable derivative. As an equivalent norm we can then take

$$\sqrt{\sum_{j=0}^{N} \int_{\mathbb{R}} |f^{(j)}(x)|^2 \, dx}.$$

It follows from the definition of the mother wavelet that $\left|\hat{\varphi}(\bullet)\right|^2 = \sum_{p\geq 1}\left|\hat{\psi}(2^p\bullet)\right|^2$ almost everywhere on $\mathbb{R}$ (see [11, p. 31]), and it is then quite easy to see that if $s \geq 0$, then $\varphi \in \mathcal{H}^s(\mathbb{R};\mathbb{C})$ if and only if $\psi \in \mathcal{H}^s(\mathbb{R};\mathbb{C})$; cf. [15, Prop. 10.2].

In the case where the wavelets φ and ψ, and then also the filter α, have compact support, the following theorem can essentially be found in [6] and [15]. The proof given below is a modification of that in [15]. For related results, see also [2], [5], and [9].

THEOREM 1. *Let $(\{V_m\}_{m\in\mathbb{Z}}, \varphi)$ be a multiresolution of $L^2(\mathbb{R};\mathbb{C})$ with filter α and mother wavelet ψ. Let $M \geq 1$ be an integer and assume that $|\bullet|^{M+1}\varphi \in L^2(\mathbb{R};\mathbb{C})$ and that*

$$\int_{\mathbb{R}} x^j \psi(x)\,\mathrm{d}x = 0, \qquad j = 0,1,2,\dots, M-1. \tag{9}$$

Let a be the function

$$a = \frac{\hat{\alpha}(\bullet)}{\left(\cos(\pi\bullet)\right)^M}.$$

Then a is continuous and periodic and the operator A: $C([0,1];\mathbb{C}) \to C([0,1];\mathbb{C})$, defined by

$$Af = \left|a\left(\frac{\bullet}{2}\right)\right|^2 f\left(\frac{\bullet}{2}\right) + \left|a\left(\frac{\bullet+1}{2}\right)\right|^2 f\left(\frac{\bullet+1}{2}\right),$$

has spectral radius $\rho \geq 1$. In particular, $\rho > 1$ if $\int_{\mathbb{R}} x^M\psi(x)\,\mathrm{d}x \neq 0$.

If $s < M - \log_4(\rho)$, then $\varphi \in \mathcal{H}^s(\mathbb{R};\mathbb{C})$ and if $\rho > 1$ and $\varphi \in \mathcal{H}^s(\mathbb{R};\mathbb{C})$, then $s < M - \log_4(\rho)$.

It is proved below that it follows from the assumption $|\bullet|^{M+1}\varphi \in L^2(\mathbb{R};\mathbb{C})$ that $|\bullet|^{M+1}\psi \in L^2(\mathbb{R};\mathbb{C})$ too, and therefore the integrals in (9) are well defined. If there is some closed subspace of $C([0,1];\mathbb{C})$ invariant under A that contains all the constants, then one sees from the fact that A is a positive operator that the spectral radius of A restricted to this subspace is ρ as well. This observation is important in the case where the parent wavelets have compact support because one can then find an invariant subspace of trigonometric polynomials and ρ is the spectral radius of a matrix.

Concerning bases, we have the following result.

THEOREM 2. *Let $s \geq 0$ and assume that $(\{V_m\}_{m\in\mathbb{Z}}, \varphi)$ is a multiresolution of $L^2(\mathbb{R};\mathbb{C})$ with mother wavelet ψ such that $\varphi \in \mathcal{H}^s(\mathbb{R};\mathbb{C})$, and $|\bullet|^N\varphi \in L^2(\mathbb{R};\mathbb{C})$ for all $N \geq 0$. Then the sets*

$$\{\psi(2^m \bullet -k)\}_{m,k\in\mathbb{Z}},$$

and

$$\{\varphi(2^{m_0} \bullet -k), \psi(2^m \bullet -k)\}_{m\geq m_0,k\in\mathbb{Z}},$$

where $m_0 \in \mathbb{Z}$ is arbitrary, are unconditional bases for $\mathcal{H}^t(\mathbb{R};\mathbb{C})$, $t \in [-s,s]$.

In [13] the corresponding assumptions are that s is an integer and that for all $j \geq 0$ one has $\sup_{x\in\mathbb{R}}(1+|x|)^j|\varphi^{(s)}(x)| < \infty$. In the case of wavelets with compact support, one can calculate exactly for which values of s we have $\varphi \in \mathcal{H}^s(\mathbb{R};\mathbb{C})$; see [6], and for these spaces we immediately get an unconditional basis as well.

As one sees from the proof, one could use a number of slightly different sets of assumptions. The ones used here are not necessarily the weakest possible, but they are reasonable and simple.

3. Proof of Theorem 1. First we give some auxiliary results. By $\mathbb{T}$ we denote $\mathbb{R}/\mathbb{Z}$, i.e., functions defined on $\mathbb{T}$ are periodic functions on $\mathbb{R}$ with period 1.

LEMMA 3. *Let $N \geq 1$ and let $(\{V_m\}_{m\in\mathbb{Z}}, \varphi)$ be a multiresolution of $L^2(\mathbb{R};\mathbb{C})$ with filter α and mother wavelet ψ such that $|\bullet|^N\varphi \in L^2(\mathbb{R};\mathbb{C})$. Then $\hat{\varphi}$, $\hat{\psi} \in \mathcal{H}^N(\mathbb{R};\mathbb{C})$ and $\hat{\alpha} \in \mathcal{H}^N(\mathbb{T};\mathbb{C})$. Moreover,*

$$\sum_{k\in\mathbb{Z}}|\hat{\varphi}(\omega+k)|^2 = \sum_{k\in\mathbb{Z}}|\hat{\psi}(\omega+k)|^2 = 1, \qquad \omega \in [0,1], \tag{10}$$

where the series converge uniformly.

Note that $\hat{\alpha} \in \mathcal{H}^N(\mathbb{T};\mathbb{C})$ means that $\hat{\alpha}$ is an $N-1$ times continuously differentiable periodic function and $\hat{\alpha}^{(N-1)}$ is absolutely continuous with a locally square integrable derivative.

Proof of Lemma 3. The fact that $\varphi \in \mathcal{H}^N(\mathbb{R};\mathbb{C})$ is a direct consequence of the moment condition and the definition of $\mathcal{H}^N(\mathbb{R};\mathbb{C})$. But from now on we use the equivalent characterization of $\mathcal{H}^N(\mathbb{R};\mathbb{C})$ in terms of square integrable derivatives.

It follows from the fact that $\{\varphi(\bullet-k)\}_{k\in\mathbb{Z}}$ and $\{\psi(\bullet-k)\}_{k\in\mathbb{Z}}$ are orthonormal sets that (10) holds for almost every $\omega \in [0,1]$; see [11, p. 31]. Because

$$\hat{\varphi}(\omega) = \hat{\varphi}(\eta) + \int_\omega^\eta \hat{\varphi}'(\xi)\,\mathrm{d}\xi, \qquad \omega,\eta \in [0,1],$$

we have

$$|\hat{\varphi}(\omega+k)|^2 \leq 2\int_0^1 |\hat{\varphi}(\eta+k)|^2\,\mathrm{d}\eta + 2\int_0^1 |\hat{\varphi}'(\xi+k)|^2\,\mathrm{d}\xi, \quad \omega \in [0,1], \quad k \in \mathbb{Z}.$$

Since $\hat{\varphi}$ and $\hat{\varphi}' \in L^2(\mathbb{R};\mathbb{C})$ we get the result that the first series in (10) converges uniformly.

It follows that for each $\omega \in [0,1]$ there exist numbers $k_\omega \in \mathbb{Z}$ and $\epsilon_\omega > 0$ such that $|\hat{\varphi}(\xi+k_\omega)| > 0$ when $|\xi-\omega| < 2\epsilon_\omega$. Since $[0,1]$ is compact we can choose finitely many of these points ω_j, $j = 1,2,\dots,n$ such that $[0,1] \subset \bigcup_{j=1}^n(\omega_j-\epsilon_{\omega_j}, \omega_j+\epsilon_{\omega_j})$. Therefore $|\hat{\varphi}(\xi+k_{\omega_j})| \geq C$ for some constant $C > 0$ when $|\xi-\omega_j| < \epsilon_{\omega_j}$ and $j = 1,\dots,n$. But $\hat{\alpha}(\xi) = \hat{\varphi}(2(\xi+k_{\omega_j}))/\hat{\varphi}(\xi+k_{\omega_j})$ by (6) and by the periodicity of $\hat{\alpha}$, and therefore we deduce that $\hat{\alpha}$ is $N-1$ times continuously differentiable, and $\hat{\alpha}^{(N-1)}$ is absolutely continuous with a square integrable derivative on the interval $(\omega_j - \epsilon_{\omega_j}, \omega_j + \epsilon_{\omega_j})$. Since $[0,1] \subset \bigcup_{j=1}^{n}(\omega_j - \epsilon_{\omega_j}, \omega_j + \epsilon_{\omega_j})$ we conclude that $\hat{\alpha} \in \mathcal{H}^N(\mathbb{T};\mathbb{C})$.

From (8) we get

$$2^N \hat{\psi}^{(N)}(2\bullet) = \sum_{j=0}^{N} \binom{N}{j} \hat{\beta}^{(j)}(\bullet)\hat{\varphi}^{(N-j)}(\bullet),$$

where $\hat{\beta}(\bullet) = -\mathrm{e}^{-2\pi \mathrm{i}\bullet}\overline{\hat{\alpha}\left(\bullet + \frac{1}{2}\right)}$. Because $\sup_{\omega\in\mathbb{R}}|\hat{\alpha}^{(j)}(\omega)| < \infty$ for $j = 0, 1, \dots, N-1$, $\int_0^1 |\hat{\alpha}^{(N)}(\omega)|^2 \, \mathrm{d}\omega < \infty$, and $\hat{\alpha}^{(N)}$ is periodic, these conclusions hold for $\hat{\beta}$ as well. Since, moreover, (10) holds and $\hat{\varphi} \in H^N(\mathbb{R};\mathbb{C})$ we see that $\int_{\mathbb{R}} |\hat{\psi}^{(N)}(\omega)|^2 \, d\omega < \infty$, and hence $\hat{\psi} \in \mathcal{H}^N(\mathbb{R};\mathbb{C})$.

Repeating the argument used above we now see that the second series in (10) converges uniformly too. □

We replace the function a defined above by $\hat{\alpha}(\bullet)(\mathrm{e}^{\pi\mathrm{i}\bullet}\cos(\pi\bullet))^{-M}$. This change will not affect the operator A, but a will now be periodic with period 1, and this is in some parts of the argument desirable (although not very essential).

We apply Lemma 3 with $N = M + 1$, and then we note that assumption (9) implies that $\hat{\psi}^{(j)}(0) = 0$ for $j = 0, 1, \dots, M-1$. By (8) and the fact that $|\hat{\varphi}(0)| = 1$ (see [11, p. 31]) it follows that we also have $\hat{\alpha}^{(j)}(\frac{1}{2}) = 0$ for $j = 0, 1, \dots, M-1$. Thus we get by Taylor's formula that

$$\hat{\alpha} = \frac{\hat{\alpha}^{(M)}(1/2)}{M!}\left(\bullet - \frac{1}{2}\right)^M + \frac{1}{M!}\int_{1/2}^{\bullet} \hat{\alpha}^{(M+1)}(\xi)(\bullet - \xi)^M \, \mathrm{d}\xi.$$

From this formula we can deduce with the aid of Hölder's inequality that a is absolutely continuous and $a' \in L^p(\mathbb{T};\mathbb{C})$, $1 \le p < 2$. Thus we can conclude that a is in fact Hölder continuous, for example, with exponent $\frac{1}{4}$, i.e.,

$$\sup_{\omega,\eta\in[0,1]} \frac{|a(\omega) - a(\eta)|}{|\omega - \eta|^{1/4}} < \infty.$$

Next we study the operator A. By changing variables and using the periodicity of $|a|$ we easily see that for all f, $g \in C(\mathbb{T};\mathbb{C})$ we have

$$\int_0^1 (Af)(\omega)g(\omega)\, \mathrm{d}\omega = \int_0^1 f(\omega)2|a(\omega)|^2 g(2\omega)\, \mathrm{d}\omega. \tag{11}$$

In particular, this implies that if we define numbers σ_m by

$$\sigma_m = 2^m \int_0^1 |\sin(\pi 2^m \omega)|^{2M} \prod_{k=0}^{m-1} |a(2^k\omega)|^2 \, \mathrm{d}\omega, \qquad m \ge 1, \tag{12}$$

then it follows from m applications of (11) that

$$\sigma_m = \int_0^1 (A^m 1)(\omega)|\sin(\pi\omega)|^{2M} \, \mathrm{d}\omega, \qquad m \ge 1. \tag{13}$$

Thus we have

$$\sigma_m \leq \|A^m\|, \qquad m \geq 1, \tag{14}$$

where $\|\bullet\|$ denotes the operator norm.

Let

$$\mu \overset{\text{def}}{=} \inf_{\omega\in\mathbb{R}} \left(|a(\omega)|^2 + \left|a\left(\omega+\tfrac{1}{2}\right)\right|^2 \right).$$

It is clear that $(A^m 1)(\omega) \geq \mu^m$, $\omega \in [0,1]$, hence $\|A^m\| \geq \mu^m$, and it follows that the spectral radius of A is at least μ. Since $|\cos(\pi\omega)| \leq 1$ for all ω we conclude from (7) and the definition of a that $\mu \geq 1$.

If $\int_{\mathbb{R}} x^M \psi(x)\,dx \neq 0$, then $\hat{\psi}^{(M)}(0) \neq 0$, and it follows that $\hat{\alpha}^{(M)}(\frac{1}{2}) \neq 0$ and therefore $a(\frac{1}{2}) \neq 0$. But $a(0) = 1$ because $\hat{\alpha}(0) = \lim_{\omega\to 0} \hat{\varphi}(2\omega)/\hat{\varphi}(\omega) = 1$ by (6), and $|\cos(\pi\omega)| < 1$ when $\omega \notin \mathbb{Z}$. Therefore it follows from (7) and the definition of a that $\mu > 1$. Thus we have established the claims about the spectral radius.

If we replace a by the standard approximation

$$a_n = \frac{1}{n+1} \int_0^1 \left(\frac{\sin((n+1)\pi t)}{\sin(\pi t)} \right)^2 a(\bullet - t)\,dt,$$

then we see that the functions a_n are uniformly Hölder continuous and bounded, and $a_n \to a$ uniformly as $n \to \infty$. We define the operator $A_n : C([0,1];\mathbb{C}) \to C([0,1];\mathbb{C})$ by

$$A_n f = \left| a_n\left(\frac{\bullet}{2}\right) \right|^2 f\left(\frac{\bullet}{2}\right) + \left| a_n\left(\frac{\bullet+1}{2}\right) \right|^2 f\left(\frac{\bullet+1}{2}\right).$$

Because $\widehat{a_n}(k) = \max\{0, 1 - (|k|/n+1)\}\hat{a}(k)$ it follows that a_n is a trigonometric polynomial. (We have defined the Fourier transform of the periodic function a to be $\hat{a} = \int_0^1 e^{-2\pi i \bullet \omega} a(\omega)\,d\omega$.) But then $|a_n|^2$ is a trigonometric polynomial as well, and the support of the Fourier transform of $|a_n|^2$ is contained in $[-2n, 2n]$. If now $f \in C([0,1];\mathbb{C})$ is a trigonometric polynomial, then

$$|a_n(\bullet)|^2 f(\bullet) = \sum_{k\in\mathbb{Z}} e^{2\pi i k \bullet} \sum_{j=-2n}^{2n} \widehat{|a_n|^2}(j) \hat{f}(k-j),$$

where the sum is actually a finite one. Then

$$A_n f = 2 \sum_{k\in\mathbb{Z}} e^{2\pi i k \bullet} \sum_{j=-2n}^{2n} \widehat{|a_n|^2}(j) \hat{f}(2k-j),$$

because the odd terms cancel. Thus we see that if the support of the Fourier transform of f is contained in $[-2n+1, 2n-1]$, then the same holds true for the support of the Fourier transform of $A_n f$. Thus A_n maps a finite-dimensional space of trigonometric polynomials into itself, and therefore there is an eigenvalue λ_n of A_n such that $|\lambda_n| = \rho_n$, where ρ_n is the spectral radius of A_n restricted to this space. We denote the corresponding eigenfunction by v_n and we normalize it so that $\|v_n\|_{C([0,1])} = 1$. We note that ρ_n is also equal to the spectral radius of A_n in the space $C([0,1];\mathbb{C})$, because $\|A_n^m\| = \|A_n^m 1\|_{C([0,1])}$, where $\|\bullet\|$ is the operator norm in any one of these spaces.

Define a new operator $B_n : C([0,1] \times [0,1]; \mathbb{C}) \to C([0,1] \times [0,1]; \mathbb{C})$ by

$$(B_n h)(\omega, \eta) = \left| a_n\left(\frac{\omega}{2}\right)\right|^2 h\left(\frac{\omega}{2}, \frac{\eta}{2}\right) + \left| a_n\left(\frac{\omega+1}{2}\right)\right|^2 h\left(\frac{\omega+1}{2}, \frac{\eta+1}{2}\right), \qquad \omega, \eta \in [0,1].$$

We can also define B in a similar way with a_n replaced by a. We note that B_n applied to a function that does not depend on its second argument gives the same result (as a function of its first argument) as A_n applied to the same function (with only one argument). Since $\|B_n^m\| = \|B_n^m 1\|_{C([0,1]^2)}$ and $\|A_n^m\| = \|A_n^m 1\|_{C([0,1])}$ we therefore conclude that $\|B_n^m\| = \|A_n^m\|$ and these operators have the same spectral radius. Moreover, $B_n \to B$ as $n \to \infty$. Define the function $g_n \in C([0,1] \times [0,1]; \mathbb{C})$ by

$$g_n(\omega, \eta) = \frac{v_n(\omega) - v_n(\eta)}{|\omega - \eta|^{1/4}}, \qquad \omega, \eta \in [0,1].$$

Then we get

$$2^{-1/4} B_n g_n - \rho_n g_n = b_n,$$

where

$$b_n(\omega, \eta) = \frac{\left(\left|a_n\left(\frac{\eta}{2}\right)\right|^2 - \left|a_n\left(\frac{\omega}{2}\right)\right|^2\right) v_n\left(\frac{\eta}{2}\right) + \left(\left|a_n\left(\frac{\eta+1}{2}\right)\right|^2 - \left|a_n\left(\frac{\omega+1}{2}\right)\right|^2\right) v_n\left(\frac{\eta+1}{2}\right)}{|\omega - \eta|^{1/4}}$$

for all ω, $\eta \in [0,1]$. Since a_n is uniformly Hölder continuous with exponent $\frac{1}{4}$, $\sup_{\eta \in [0,1]} |v_n(\eta)| = 1$, $\rho_n \to \rho$, and $B_n \to B$ as $n \to \infty$, we conclude that

$$\sup_{n \geq 1} \sup_{\omega, \eta \in [0,1]} |g_n(\omega, \eta)| < \infty.$$

But this means that the functions v_n are uniformly Hölder continuous, in particular, equicontinuous, and we may pass to the limit and get a nontrivial function $v \in C([0,1]; \mathbb{C})$ such that $Av = \lambda v$, where $|\lambda| = \rho$. But then we have by (13), because $\sup_{\omega \in [0,1]} |v(\omega)| = 1$,

$$\begin{aligned} \rho^m \int_0^1 |v(\omega)|^2 |\sin(\pi\omega)|^{2M} \, d\omega &= \left| \int_0^1 (A^m v)(\omega) \overline{v(\omega)} |\sin(\pi\omega)|^{2M} \, d\omega \right| \\ &\leq \int_0^1 (A^m 1)(\omega) |\sin(\pi\omega)|^{2M} \, d\omega = \sigma_m. \end{aligned}$$

Combining this result with (14) and the fact that $\lim_{m \to \infty} \|A^m\|^{1/m} = \rho$, we conclude that

$$\sum_{m=1}^{\infty} 4^{-pm} \sigma_m < \infty \quad \text{if and only if } p > \log_4(\rho). \tag{15}$$

Let us define the function q by

$$q = \frac{(\pi \bullet)^M \hat{\varphi}(\bullet)}{\left(e^{\pi i \bullet} \sin(\pi \bullet)\right)^M \hat{\varphi}(0)}.$$

Using the same kind of argument that was used when proving that a was continuous, we can prove that q is continuous as well. The important point, however, is that since $\prod_{k=1}^{\infty}\cos(\pi 2^{-k}\bullet) = \sin(\pi\bullet)/(\pi\bullet)$ and $\hat{\varphi} = \hat{\varphi}(0)\prod_{k=1}^{\infty}\hat{\alpha}(2^{-k}\bullet)$ (see [11, p. 34]), it follows that

$$q = \prod_{k=1}^{\infty} a(2^{-k}\bullet). \tag{16}$$

(Recall that we introduced a factor $e^{\pi i \bullet}$ in the definition of a for technical reasons.)

By Lemma 3 there exists an integer J such that $\sum_{j=-J}^{J-1}|\hat{\varphi}(\omega+j)|^2 \geq \frac{1}{2}$ for all $\omega \in [0,1]$. Since $|q(\omega)| \geq |\hat{\varphi}(\omega)|$, $\omega \in \mathbb{R}$, we have for some constant C,

$$\frac{1}{2} \leq \sum_{j=-J}^{J-1} |q(\omega+j)|^2 \leq C, \qquad \omega \in [0,1]. \tag{17}$$

Let $p > 0$. It follows from the definition of q that

$$\varphi \in \mathcal{H}^{M-p}(\mathbb{R};\mathbb{C}) \quad \text{if and only if} \quad \int_{\mathbb{R}} (1+\omega^2)^{-p}|q(\omega)|^2\big|\sin(\pi\omega)\big|^{2M}\,d\omega < \infty.$$

Because $p > 0$ we get after integrating by parts that

$$\begin{aligned}&\int_{\mathbb{R}} (1+\omega^2)^{-p}|q(\omega)|^2\big|\sin(\pi\omega)\big|^{2M}\,d\omega \\ &\quad = \int_0^{\infty} 2p\omega(1+\omega^2)^{-p-1}\int_{-\omega}^{\omega} |q(\eta)|^2\big|\sin(\pi\eta)\big|^{2M}\,d\eta\,d\omega,\end{aligned}$$

where equality holds in the case where one of the integrals diverges too. There are positive constants C_1 and C_2 such that $C_1 \leq 4^{pm}\int_{2^{m-1}J}^{2^m J} 2p\omega(1+\omega^2)^{-p-1}\,d\omega \leq C_2$, and hence we see that

$$\varphi \in \mathcal{H}^{M-p}(\mathbb{R};\mathbb{C}) \quad \text{if and only if} \quad \sum_{m=1}^{\infty} 4^{-pm}\int_{-2^m J}^{2^m J} |\sin(\pi\omega)|^{2M}|q(\omega)|^2\,d\omega < \infty. \tag{18}$$

Using (16), changing variables, and invoking the periodicity of $|a|$ we get

$$\begin{aligned}&\int_{-2^m J}^{2^m J} |\sin(\pi\omega)|^{2M}|q(\omega)|^2\,d\omega \\ &\quad = \int_{-2^m J}^{2^m J} |\sin(\pi\omega)|^{2M}\prod_{k=1}^{m}|a(2^{-k}\omega)|^2|q(2^{-m}\omega)|^2\,d\omega \\ &\quad = 2^m\int_{-J}^{J} |\sin(\pi 2^m\omega)|^{2M}\prod_{k=0}^{m-1}|a(2^k\omega)|^2|q(\omega)|^2\,d\omega \\ &\quad = 2^m\int_0^1 |\sin(\pi 2^m\omega)|^{2M}\prod_{k=0}^{m-1}|a(2^k\omega)|^2\sum_{j=-J}^{J-1}|q(\omega+j)|^2\,d\omega.\end{aligned}$$

If we combine this result with (12), (15), (17), and (18), then we see that when $p > 0$ we have $\varphi \in \mathcal{H}^{M-p}(\mathbb{R};\mathbb{C})$ if and only if $p > \log_4(\rho)$. But this is exactly the claim of the theorem, and the proof is completed. □

4. Proof of Theorem 2. We use the following result on projections onto spaces spanned by translations of one function.

LEMMA 4. *Let $\phi \in L^2(\mathbb{R};\mathbb{C})$ be such that $\{\phi(\bullet - k)\}_{k\in\mathbb{Z}}$ is an orthonormal set in $L^2(\mathbb{R};\mathbb{C})$. Let $m \in \mathbb{Z}$ and denote by P_m the orthogonal projection onto the closed subspace of $L^2(\mathbb{R};\mathbb{C})$ spanned by $\{\phi(2^m \bullet -k)\}_{k\in\mathbb{Z}}$. Then, for every $f \in L^2(\mathbb{R};\mathbb{C})$,*

$$\widehat{P_m f} = \hat{\phi}(2^{-m}\bullet)g_m(\bullet), \tag{19}$$

where

$$g_m = \sum_{j\in\mathbb{Z}} \hat{f}(\bullet + 2^m j)\overline{\hat{\phi}(2^{-m}\bullet + j)}. \tag{20}$$

Moreover,

$$\|P_m f\|^2_{L^2(\mathbb{R})} = \int_{-2^{m-1}}^{2^{m-1}} |g_m(\omega)|^2\,\mathrm{d}\omega,$$

and for every $t \in \mathbb{R}$,

$$\|P_m f\|^2_{\mathcal{H}^t(\mathbb{R})} = \int_{-2^{m-1}}^{2^{m-1}} \sum_{k\in\mathbb{Z}} (1 + |\omega + 2^m k|^2)^t \left|\hat{\phi}(2^{-m}\omega + k)\right|^2 |g_m(\omega)|^2\,\mathrm{d}\omega.$$

Proof of Lemma 4. First we derive a formula for $\widehat{P_m f}$. We have

$$P_m f = \sum_{k\in\mathbb{Z}} 2^m \langle f, \phi(2^m \bullet -k)\rangle \phi(2^m \bullet -k),$$

where the series converges in $L^2(\mathbb{R};\mathbb{C})$ and $\langle \bullet, \bullet\rangle$ denotes the inner product in $L^2(\mathbb{R};\mathbb{C})$. Now the Fourier transform of $\phi(2^m \bullet -k)$ is $2^{-m}\mathrm{e}^{-2\pi \mathrm{i} 2^{-m}k\bullet}\hat{\phi}(2^{-m}\bullet)$, and therefore we have

$$\widehat{P_m f} = \hat{\phi}(2^{-m}\bullet)\sum_{k\in\mathbb{Z}} \mathrm{e}^{-2\pi\mathrm{i}2^{-m}k\bullet}\langle f, \phi(2^m \bullet -k)\rangle.$$

Using Plancherel's theorem we get

$$\begin{aligned}\langle f, \phi(2^m \bullet -k)\rangle &= \int_{\mathbb{R}} \hat{f}(\omega)2^{-m}\mathrm{e}^{2\pi\mathrm{i}2^{-m}k\omega}\overline{\hat{\phi}(2^{-m}\omega)}\,\mathrm{d}\omega \\ &= \int_0^1 \mathrm{e}^{2\pi\mathrm{i}k\omega}\sum_{j\in\mathbb{Z}} \hat{f}(2^m(\omega + j))\overline{\hat{\phi}(\omega + j)}\,\mathrm{d}\omega,\end{aligned}$$

and we see that if we define the function g_m by (20), then

$$\langle f, \phi(2^m \bullet -k)\rangle = \widehat{g_m(2^m\bullet)}(-k). \tag{21}$$

Now we combine this result with the Fourier inversion theorem for square integrable periodic functions and conclude that (19) holds.

We have $\|P_m f\|^2_{L^2(\mathbb{R})} = 2^m \sum_{k\in\mathbb{Z}} |\langle f, \phi(2^m \bullet -k)\rangle|^2$ because $\{2^{m/2}\phi(2^m \bullet -k)\}_{k\in\mathbb{Z}}$ is an orthonormal set, and therefore it follows from (21) and Plancherel's theorem that

$$\begin{aligned}\|P_m f\|^2_{L^2(\mathbb{R})} &= 2^m \sum_{k\in\mathbb{Z}} |\widehat{g_m(2^m\bullet)}(-k)|^2 \\ &= 2^m \int_{-1/2}^{1/2} |g_m(2^m\omega)|^2\,\mathrm{d}\omega = \int_{-2^{m-1}}^{2^{m-1}} |g_m(\omega)|^2\,\mathrm{d}\omega.\end{aligned}$$

The last claim is an immediate consequence of the fact that g_m is periodic with period 2^m. This completes the proof. □

It follows from the assumptions and Lemma 3 that $\hat{\varphi}$, $\hat{\psi} \in C^\infty(\mathbb{R};\mathbb{C})$ and that $\hat{\alpha} \in C^\infty(\mathbb{T};\mathbb{C})$. If there is an integer $M \geq 1$ such that (9) holds but $\int_{\mathbb{R}} x^M \psi(x)\,\mathrm{d}x \neq 0$, then it follows from Theorem 1 that $s < M - \log_4(\rho)$. If no such integer M exists, then (9) holds for every $M \geq 1$ and we can fix $M > s$ to be an arbitrary integer. But then we have again $s < M - \log_4(\rho)$ either because $\rho > 1$, and we can apply Theorem 1, or because $\rho = 1$ and $s < M = M - \log_4(\rho)$. Thus we see that there exists a number $\delta > 0$ such that $s + 2\delta < M - \log_4(\rho)$, that is, $\varphi \in \mathcal{H}^{s+2\delta}(\mathbb{R};\mathbb{C})$ by Theorem 1. Moreover, we get that

$$\hat{\psi}^{(j)}(0) = 0, \qquad j = 0, 1, \dots, \lfloor s+\delta \rfloor. \tag{22}$$

Next we show that

$$\begin{aligned} &\sup_{\omega\in\mathbb{R}} \sum_{k\in\mathbb{Z}} \left(1 + |\omega + k|^2\right)^{s+\delta} |\hat{\varphi}(\omega + k)|^2 < \infty, \\ &\sup_{\omega\in\mathbb{R}} \sum_{k\in\mathbb{Z}} \left(1 + |\omega + k|^2\right)^{s+\delta} |\hat{\psi}(\omega + k)|^2 < \infty. \end{aligned} \tag{23}$$

Since $\sup_{\omega\in\mathbb{R}} |\hat{\alpha}(\omega)| \leq 1$ by (7) and therefore $|\hat{\psi}(2\omega)| \leq |\hat{\varphi}(\omega)|$, $\omega \in \mathbb{R}$, by (6), we see that the second claim is a consequence of the first; so we will only prove that one.

Choose an integer m so large that $m \geq 1 + s/(2\delta)$. It follows from one of the basic Sobolev inequalities (see, e.g., [7, p. 27]) that there is a constant C_1 such that for all $\epsilon \in (0,1)$ and all $k \in \mathbb{Z}$,

$$\sup_{\omega\in[0,1]} |\hat{\varphi}(\omega + k)|^2 \leq C_1\epsilon \int_0^1 |\hat{\varphi}^{(m)}(\xi + k)|^2\,\mathrm{d}\xi + C_1\epsilon^{-1/(2m-1)} \int_0^1 |\hat{\varphi}(\xi + k)|^2\,\mathrm{d}\xi.$$

Choose $\epsilon = (1 + k^2)^{-(s+\delta)}$ and note that $\epsilon^{-1/(2m-1)}(1 + k^2)^{s+\delta} \leq (1 + k^2)^{s+2\delta}$. We conclude that

$$\begin{aligned} &\sup_{\omega\in[0,1]} (1 + k^2)^{s+\delta} |\hat{\varphi}(\omega + k)|^2 \\ &\qquad \leq C_1 \int_0^1 |\hat{\varphi}^{(m)}(\xi + k)|^2\,\mathrm{d}\xi + C_1 \int_0^1 (1 + k^2)^{s+2\delta} |\hat{\varphi}(\xi + k)|^2\,\mathrm{d}\xi. \end{aligned}$$

Now we have $\sum_{k\in\mathbb{Z}} \int_0^1 |\hat{\varphi}^{(m)}(\xi + k)|^2\,\mathrm{d}\xi < \infty$ because $|\bullet|^m \varphi(\bullet) \in L^2(\mathbb{R};\mathbb{C})$ and $\sum_{k\in\mathbb{Z}} \int_0^1 (1 + |\xi + k|^2)^{s+2\delta} |\hat{\varphi}(\xi + k)|^2\,\mathrm{d}\xi < \infty$ because $\varphi \in \mathcal{H}^{s+2\delta}(\mathbb{R};\mathbb{C})$, and therefore we get the desired conclusion.

Denote the orthogonal projection onto the space spanned by $\{\varphi(2^m \bullet - k)\}_{k\in\mathbb{Z}}$ by P_m and the one onto the space spanned by $\{\psi(2^m \bullet - k)\}_{k\in\mathbb{Z}}$ by Q_m. We claim that there exists a constant C_2 such that for every $m \geq 0$, $f \in L^2(\mathbb{R};\mathbb{C})$, and every t with $|t| \leq s + \delta$, we have

$$\begin{aligned} \|P_0 f\|_{\mathcal{H}^t(\mathbb{R})} &\leq C_2 \|P_0 f\|_{L^2(\mathbb{R})}, \\ \|Q_m f\|_{\mathcal{H}^t(\mathbb{R})} &\leq C_2 2^{mt} \|Q_m f\|_{L^2(\mathbb{R})}. \end{aligned} \tag{24}$$

If $t \geq 0$ these inequalities follow immediately from Lemma 4 and (23). If $t < 0$ the first inequality is trivially true, and in order to get the second one we note that by

(22) there exists a constant C_3 such that $|\hat{\psi}(\eta)| \leq C_3|\eta|^{s+\delta}$ when $|\eta| \leq 1$. Thus we see from some easy estimates and (10) that

$$\begin{aligned}
&\sup_{|\omega|\leq 2^{m-1}} \sum_{k\in\mathbf{Z}} \left(1+|\omega+k2^m|^2\right)^t \left|\hat{\psi}(2^{-m}\omega+k)\right|^2 \\
&\leq \sup_{|\omega|\leq 2^{m-1}} (1+|\omega|^2)^t |\hat{\psi}(2^{-m}\omega)|^2 \\
&\quad + \sup_{|\omega|\leq 2^{m-1}} \sum_{k\in\mathbf{Z}\setminus\{0\}} \left(1+|\omega+k2^m|^2\right)^t \left|\hat{\psi}(2^{-m}\omega+k)\right|^2 \\
&\leq C_3 2^{2mt}|\omega|^{-2t}(1+|\omega|^2)^t + 2^{2mt-2t}\sum_{k\in\mathbf{Z}}\left|\hat{\psi}(2^{-m}\omega+k)\right|^2 \\
&\leq 2^{2mt}(C_3+4^{-t}),
\end{aligned}$$

and the last claim follows from Lemma 4 as well.

The rest of the proof now follows that in [13]. Let $f \in L^2(\mathbb{R};\mathbb{C})$ and $t \in [-s,s]$ be arbitrary. Since $f = P_0 f + \sum_{m=0}^{\infty} Q_m f$ and $|\langle g,h\rangle_{\mathcal{H}^t(\mathbb{R})}| \leq \|g\|_{\mathcal{H}^{t-\delta}(\mathbb{R})}\|h\|_{\mathcal{H}^{t+\delta}(\mathbb{R})}$ by Hölder's inequality, we get from (24) that

$$\begin{aligned}
\|f\|^2_{\mathcal{H}^t(\mathbb{R})} &\leq 2\|P_0 f\|^2_{\mathcal{H}^t(\mathbb{R})} + 2\sum_{m=0}^{\infty}\|Q_m f\|^2_{\mathcal{H}^t(\mathbb{R})} + 4\sum_{m=1}^{\infty}\sum_{k=0}^{m-1}\left|\langle Q_m f, Q_k f\rangle_{\mathcal{H}^t(\mathbb{R})}\right| \\
&\leq 2C_2^2\left(\|P_0 f\|^2_{L^2(\mathbb{R})} + \sum_{m=0}^{\infty} 4^{mt}\|Q_m f\|^2_{L^2(\mathbb{R})}\right) \\
&\quad + 4\sum_{m=1}^{\infty}\sum_{k=0}^{m-1}\|Q_m f\|_{\mathcal{H}^{t-\delta}(\mathbb{R})}\|Q_k f\|_{\mathcal{H}^{t+\delta}(\mathbb{R})} \\
&\leq 4C_2^2\Bigg(\|P_0 f\|^2_{L^2(\mathbb{R})} + \sum_{m=0}^{\infty} 4^{mt}\|Q_m f\|^2_{L^2(\mathbb{R})} \\
&\qquad + \sum_{m=1}^{\infty} 2^{mt}\|Q_m f\|_{L^2(\mathbb{R})} \sum_{k=0}^{m-1} 2^{-\delta(m-k)}2^{kt}\|Q_k f\|_{L^2(\mathbb{R})}\Bigg).
\end{aligned}$$

Next we observe that $\sum_{k=1}^{\infty} 2^{-\delta k} < \infty$ is an upper bound for the the norm of the mapping $\eta \in \ell^2(\mathbb{N};\mathbb{R}) \to \sum_{k=0}^{\bullet-1} 2^{-\delta(\bullet-k)}\eta(k) \in \ell^2(\mathbb{N};\mathbb{R})$, and hence there exists a constant C such that

$$\|f\|^2_{\mathcal{H}^t(\mathbb{R})} \leq C\left(\|P_0 f\|^2_{L^2(\mathbb{R})} + \sum_{m=0}^{\infty} 4^{mt}\|Q_m f\|^2_{L^2(\mathbb{R})}\right), \qquad t \in [-s,s]. \tag{25}$$

In order to get a lower bound for $\|f\|^2_{\mathcal{H}^t(\mathbb{R})}$ we choose an arbitrary integer $M > 0$ and take

$$g = P_0 f + \sum_{m=0}^{M} 4^{mt} Q_m f.$$

Then

$$|\langle f,g\rangle_{L^2(\mathbb{R})}| = \|P_0 f\|^2_{L^2(\mathbb{R})} + \sum_{m=0}^{M} 4^{mt}\|Q_m f\|^2_{L^2(\mathbb{R})}. \tag{26}$$

But on the other hand we have by Hölder's inequality and (25) that

$$\begin{aligned}|\langle f,g\rangle_{L^2(\mathbb{R})}| &\le \|f\|_{\mathcal{H}^t(\mathbb{R})}\|g\|_{\mathcal{H}^{-t}(\mathbb{R})}\\ &\le \|f\|_{\mathcal{H}^t(\mathbb{R})}\sqrt{C}\sqrt{\|P_0f\|^2_{L^2(\mathbb{R})}+\sum_{m=0}^{M}4^{-mt}4^{2mt}\|Q_mf\|^2_{L^2(\mathbb{R})}}.\end{aligned}$$

If we combine this result with (26) and let $M\to\infty$, then we get

$$\text{(27)}\qquad \frac{1}{C}\left(\|P_0f\|^2_{L^2(\mathbb{R})}+\sum_{m=0}^{\infty}4^{mt}\|Q_mf\|^2_{L^2(\mathbb{R})}\right)\le\|f\|^2_{\mathcal{H}^t(\mathbb{R})},\qquad t\in[-s,s].$$

To complete the proof, let $P_{0,k}$ and $Q_{m,k}$ be the orthogonal (in $L^2(\mathbb{R};\mathbb{C})$) projections onto the spaces spanned by $\varphi(\bullet-k)$ and $\psi(2^m\bullet-k)$, respectively. By the orthogonality of the projections and the fact that $P_{m+1}=P_m+Q_m$ (see [11, p. 31]) we have

$$\text{(28)}\qquad \begin{aligned}\|P_0f\|^2_{L^2(\mathbb{R})}&=\sum_{k\in\mathbb{Z}}\|P_{0,k}f\|^2_{L^2(\mathbb{R})}=\sum_{m=-\infty}^{-1}\sum_{k\in\mathbb{Z}}\|Q_{m,k}f\|^2_{L^2(\mathbb{R})},\\ \|Q_mf\|^2_{L^2(\mathbb{R})}&=\sum_{k\in\mathbb{Z}}\|Q_{m,k}f\|^2_{L^2(\mathbb{R})}.\end{aligned}$$

Since $P_{0,k}f=\langle f,\varphi(\bullet-k)\rangle\varphi(\bullet-k)$ and $Q_{m,k}f=2^m\langle f,\psi(2^m\bullet-k)\rangle\psi(2^m\bullet-k)$, we get the desired conclusion in the case $m_0=0$ from (25), (27), (28), and [10, Thm. 7.1].

If $m_0\neq 0$, we can use essentially the same argument as in the case where $m_0=0$. This completes the proof. □

REFERENCES

[1] A. COHEN, *Ondelettes, analyses multirésolutions et filtres miroirs en quadrature*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 439–459.

[2] J. P. CONZE, *Sur la régularité des solutions d'une équation fonctionelle*, Laboratoire de probabilitiés, Université de Rennes 1, preprint.

[3] I. DAUBECHIES, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[4] ———, *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[5] I. DAUBECHIES AND J. C. LAGARIAS, *Two-scale difference equations* I. *Existence and global regularity of solutions*, SIAM J. Math. Anal., 22 (1991), pp. 1388–1410.

[6] T. EIROLA, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.

[7] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

[8] C. E. HEIL AND D. F. WALNUT, *Continuous and discrete wavelet transforms*, SIAM Rev., 31 (1989), pp. 628–666.

[9] L. HERVE, *Constructions et régularité des fonctions d'échelle*, Laboratoire de probabilitiés, Université de Rennes 1, preprint.

[10] R. C. JAMES, *Bases in Banach spaces*, Amer. Math. Monthly, 89 (1982), pp. 625–640.

[11] P. G. LEMARIÉ, *Analyse multi-echelles et ondelettes a support compact*, in Les Ondelettes en 1989, P. G. Lemarié, ed., Lecture Notes in Math. 1438, Springer-Verlag, Berlin, 1990, pp. 26–38.

[12] S. G. MALLAT, *Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$*, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.

[13] Y. MEYER, *Ondelettes et Opérateurs* I, Hermann, Paris, 1990.

[14] G. STRANG, *Wavelets and dilation equations: a brief introduction*, SIAM Rev., 31 (1989), pp. 614– 627.

[15] L. F. VILLEMOES, *Energy moments in time and frequency for two-scale difference equation solutions and wavelets*, SIAM J. Math. Anal., 23 (1992), pp. 1519–1543.

SIAM J. MATH. ANAL.
Vol. 24, No. 4, pp. 1043–1066, July 1993

011

LEAST SQUARES APPROXIMATION BY RADIAL FUNCTIONS*

E. QUAK[†], N. SIVAKUMAR[†], AND J. D. WARD[†][‡]

This paper is dedicated to the memory of Professor Lothar Collatz.

Abstract. This paper is concerned with the study of continuous least squares approximation on a bounded domain in $\mathbb{R}^s$ by certain classes of radial functions. The approximating subspace is spanned by translates $F(\cdot - x_j)$ of a given radial function F, where the (distinct) "centers" $\{x_j\}_{j=1}^N$ are allowed to be scattered. The main result gives quantitative estimates for the Euclidean norms of the inverses of these least squares matrices. In general, the estimates involve the dimension of the ambient space, the minimal separation distance between the centers, the number of centers, and of course the function itself. However, if F is the scaled Gaussian, it is possible to dispense with the dependence on the number of centers. Also established along the way are results involving radial interpolation matrices where the interpolation points are small perturbations of the centers. These results are perhaps of independent interest as previous interpolation results had been obtained only for interpolation at the centers.

Key words. conditionally negative definite, completely monotone, interpolation, least squares, radial

AMS subject classifications. 41A05, 41A63

1. Introduction.

The purpose of this paper is to investigate the problem of continuous least squares approximation by translates of radial functions from certain classes. Let F denote a radial function on $\mathbb{R}^s$ and $X := \{x_i\}_{i=1}^N$ a set of distinct points (called centers) in a bounded, closed domain $\Omega \subset \mathbb{R}^s$. Define $h_i(x) := F(x - x_i)$ and $S_X :=$ span$\{h_i\colon 1 \le i \le N\}$. In recent years, considerable attention has been paid to the problem of interpolation from the space S_X, i.e., finding suitable classes of radial functions F, so that for given data $d_1, \ldots, d_N \in \mathbb{R}$, there exists a unique function $h \in S_X$ satisfying the interpolation conditions $h(x_i) = d_i$, $i = 1, \ldots, N$; or, equivalently, the interpolation matrix $(h_i(x_j))_{i,j=1}^N$ is nonsingular. If F is a conditionally positive definite radial function of order zero or a conditionally negative radial function of order one (see §2 for the appropriate definitions), then it is known that the associated interpolation matrix is invertible [M], [MN], [S1]. Quantitative estimates for the Euclidean norms of the inverses of these matrices have been derived in [B], [NW1], [NW2], [Su2], [Ba].

One consequence of the invertibility of the aforementioned interpolation matrix is the linear independence of the functions h_i, i.e., the space S_X is N-dimensional. Our aim in this paper is to study certain aspects of the problem of continuous least squares approximation from the space S_X, where the underlying radial function F belongs to either of the two classes mentioned above.

More precisely, we are interested in the following issue: let $X = \{x_1, \ldots, x_N\}$ be a set of centers such that the minimal separation distance $2q := \min_{j \ne \ell} \|x_j - x_\ell\| > 0$. We call a closed and bounded domain $\Omega \subset \mathbb{R}^s$ *admissible with respect to X*, if for

*Received by the editors June 9, 1991; accepted for publication (in revised form) October 7, 1992.

[†]Center for Approximation Theory, Department of Mathematics, Texas A&M University, College Station, Texas 77843-3368.

[‡]The work of this author was supported by National Science Foundation grant DMS-8901345.

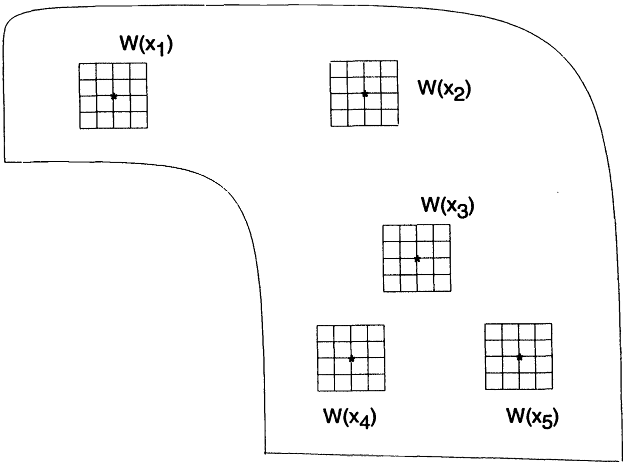


FIG. 1

each x_i, the s-dimensional cube $W(x_i)$ centered at x_i, and with side length $2qs^{-1/2}$, is contained in Ω (see Fig. 1).

Suppose that $f \in C(\Omega)$. The coefficients (with respect to the basis $\{h_i\}$) of the best least squares approximant to f from S_X are the components of the solution vector α to the matrix equation $A_X\alpha = z$, where A_X is the $N \times N$ matrix whose (i,j)th entry is $\langle h_i, h_j\rangle$, with $\langle\ ,\ \rangle$ being the standard inner product on $L^2(\Omega)$. The components of the N-vector z are $\langle h_i, f\rangle$, $i = 1,\dots,N$ (see, for instance, [de B], [C]). The primary intent of this paper (Theorem 4.6, Corollary 4.7, Propositions 4.8, and 4.10) is to give quantitative estimates for $\|A_X^{-1}\|_2$, i.e., the Euclidean norm of the inverse of the matrix A_X for continuous least squares approximation over admissible domains and associated with conditionally positive definite radial functions of order zero and conditionally negative definite radial functions of order one.

The present paper is restricted to the continuous least squares problem because its theoretical aspects are more tractable than those of its discrete counterpart. Yet, it is our belief that the basic principles of our arguments (notably the approach outlined in the following paragraph) carry over to the case of discrete least squares approximation. Thus, our paper could serve as a preliminary step towards the eventual treatment of the discrete problem.

Our approach here is to relate the problem of continuous least squares approximation from S_X to that of interpolation from S_X. However, this interpolation problem is not the standard one mentioned earlier (i.e., where the centers and the interpolation points coincide), but rather one where the interpolation points are small perturbations of the centers. Our results pertaining to this latter interpolation problem (e.g., Theorem 3.2, Theorem 3.6, and Proposition 3.11), though auxiliary for our purposes here, are, in fact, new and may also be of some independent interest.

In general, the estimates we obtain for $\|A_X^{-1}\|_2$ depend on the dimension s of the underlying space, the minimal separation distance $2q$ between the centers, the number of centers N, and of course the function F itself. The only restriction placed on Ω is that it be admissible with respect to the given set X of centers. As is to be expected, specific information regarding the roles of q and N can only be ascertained by a case by case analysis of the various radial functions of interest. However, our methods are sufficiently general in scope to allow readers to carry out such a careful analysis for functions of their choice. For our part, we have chosen four functions for purposes of illustration: the "absolute value" function, the Hardy multiquadric, the inverse multiquadric and the scaled Gaussian. It is shown (Proposition 4.8) that for the first three functions, the influence of N on the estimates for $\|A_X^{-1}\|_2$ is of at most polynomial growth, whereas the influence of the minimal separation distance essentially stems from the role of q in the interpolation problem described in the previous paragraph. On the other hand, the Gaussian differs from the other three functions in that it is possible to bound the corresponding $\|A_X^{-1}\|_2$ by a quantity independent of N (Proposition 4.10).

With regard to the conditioning of the least squares matrices considered in this paper, the crucial parameters that govern our estimates are the number of centers N and the separation parameter q. Usually, in many applications of least squares approximation, the number N is kept relatively small. Therefore, the negative impact of N, if any, on the conditioning should not be too severe (in the case of the scaled Gaussian, the number N plays no role at all). As we reduce questions concerning least squares to those involving interpolation, it stands to reason that our estimates for least squares reflect the good (or bad) behaviour of the corresponding interpolation matrices. It turns out that these matrices are very sensitive to the separation parameter q. However, the least squares problem allows us the freedom of choosing the centers in an appropriate fashion; e.g., in certain instances, it may be possible to keep the minimal separation distance quite large (especially when N is reasonably small). Furthermore, there are also other ways of improving the conditioning of the least squares problem. Often, the radial functions in question offer additional parameters (whose influence can be clearly quantified) that can be used to counter the effects of small separation distances, e.g., the constant c in the Hardy multiquadric or the scaling parameter in the Gaussian. A judicious choice of these parameters effectively leads to a new basis which, while being significantly better conditioned than the old one, can still be analyzed along the same lines as the original basis itself (Remarks 4.9(ii) and 4.11).

Certain important aspects of least squares approximation are not addressed in this paper. In particular, we do not discuss the rate of approximation as the dimension N of the approximating subspace S_X increases. Results of this nature have been obtained recently in [BDR].

We close this section with an outline of the paper: in §2, we discuss necessary preliminaries concerning conditionally negative (positive) definite radial functions. The third section, which is rather technical in nature, deals with the issue of obtaining estimates for certain interpolation matrices where the interpolation points and centers do not coincide. The main results are presented in §4. The paper concludes with §5 where sharper results are obtained in some specific instances. The optimality of one such estimate is also demonstrated therein.

2. Background and preliminaries.

We begin with a discussion of a class of functions which is well known for its multivariate interpolation properties [D], [GV], [P].

DEFINITION 2.1. A continuous function F: $\mathbb{R}^s \to \mathbb{C}$ is said to be conditionally negative (positive) definite of order m if for every finite set $\{x_j\}_{j=1}^N$ of distinct points in $\mathbb{R}^s$, and for every set of complex numbers $\{c_j\}_{j=1}^N$ satisfying

$$\sum_{j=1}^{N} c_j q(x_j) = 0$$

for every $q \in \Pi_{m-1}$ (the space of s-variate polynomials of total degree at most $m-1$), we have

$$\sum_{j,k=1}^{N} \bar{c}_j c_k F(x_j - x_k) \le 0 \qquad (\ge 0).$$

This class of conditionally negative (positive) definite functions of order m (on $\mathbb{R}^s$) will be denoted by $N_m^s(P_m^s)$.

DEFINITION 2.2. A continuous function g: $\mathbb{R}^+ \to \mathbb{R}$ is said to be a conditionally negative (positive) definite *radial* function of order m if $g \circ (\|\cdot\|)$ is in $N_m^s(P_m^s)$. (Henceforth, $\|\cdot\|$ will denote the standard Euclidean (ℓ_2) norm in $\mathbb{R}^s$.) We denote the set of all such functions by $RN_m^s(RP_m^s)$. The class RN_m^s includes those functions g which are continuous on $[0,\infty)$ and for which $(-1)^{m+1}d^m/d\sigma^m g(\sqrt{\sigma})$ is completely monotonic on $(0,\infty)$; i.e.,

$$(-1)^{m+1}\frac{d^m}{d\sigma^m} g(\sqrt{\sigma}) = \int_0^\infty e^{-\sigma t} d\mu(t),$$

where $d\mu(t)$ is some nonnegative measure on $[0,\infty)$. This latter class of functions is denoted by RN_m^∞. We also define the class RP_m^∞ by requiring that f belong to RP_m^∞ precisely when $-f$ belongs to RN_m^∞. Now suppose that F: $\mathbb{R}^s \to \mathbb{R}$ is continuous and that it is *radially symmetric*, i.e., $F(x) = F(y)$ if $\|x\| = \|y\|$, x, $y \in \mathbb{R}^s$. It is clear that F may be identified with the following function g_F: $\mathbb{R}^+ \to \mathbb{R}$, given by

$$g_F(r) = F(x), \quad \text{where } \|x\| = r.$$

Consequently, we will indulge in a slight abuse of notation and say that a function F: $\mathbb{R}^s \to \mathbb{R}$ belongs to $RN_m^\infty(RP_m^\infty)$ if F is continuous, radially symmetric, and its associated function g_F (as defined above) is in $RN_m^\infty(RP_m^\infty)$. Our emphasis in this paper will be on functions F: $\mathbb{R}^s \to \mathbb{R}$ that belong to RP_0^∞ or RN_1^∞. Although this may seem quite specialized, these two cases indeed cover many of the radial functions of interest [P]. Functions in RP_0^∞ and RN_1^∞ possess useful representations which we intend to exploit. Indeed, if $F \in RP_0^\infty$, then [S2]

$$F(x) = \int_0^\infty e^{-\|x\|^2 t} d\mu(t), \qquad x \in \mathbb{R}^s, \tag{2.1}$$

where $d\mu(t)$ is a positive measure satisfying the conditions

$$\int_0^1 d\mu(t) < \infty; \qquad \int_1^\infty e^{-t} d\mu(t) < \infty.$$

On the other hand, if $F \in RN_1^\infty$, then F may be realized as [M], [NW2], [S2], [Sul]

$$F(x) = F(0) + \int_0^\infty \frac{1 - e^{-\|x\|^2 t}}{t} d\mu(t), \qquad x \in \mathbb{R}^s, \tag{2.2}$$

where $d\mu(t)$ is a positive measure such that

$$\int_0^1 d\mu(t) < \infty; \qquad \int_1^\infty \frac{d\mu(t)}{t} < \infty.$$

While the results obtained in this paper are quite general, we wish to intersperse them with certain specific illustrations. For such purposes, we shall use the following four functions: $F_1(x) := \|x\|$, $F_2(x) := (1+\|x\|^2)^{1/2}$, $F_3(x) := (1+\|x\|^2)^{-1/2}$, and $F_4(x) := e^{-\rho\|x\|^2}$ $(\rho > 0)$. It is known that $F_1 \in RN_1^\infty$, and is represented by the measure $d\mu_1(t) := \frac{1}{2}(\pi t)^{-1/2}dt$. The Hardy multiquadric F_2 belongs to RN_1^∞ and has the representing measure $d\mu_2(t) := \frac{1}{2}e^{-t}(\pi t)^{-1/2}dt$. Both the inverse multiquadric F_3 and the scaled Gaussian F_4 belong to RP_0^∞; while F_3 is represented by the measure $2d\mu_2(t)$, the measure $d\mu_4(t) := \delta(t-\rho)dt$ (i.e., point evaluation at ρ) represents F_4. The reader, if pressed to do so, will no doubt be able to carry out similar analyses for other radial functions as well.

3. Concerning some quadratic forms. In the following, we assume that we are given two sequences, $\{x_j\}_{j=1}^N$, $\{y_j\}_{j=1}^N$, of points in $\mathbb{R}^s$ such that

$$(3.1)\qquad \begin{array}{ll} \text{(i)} & x_j \neq x_\ell, y_j \neq y_\ell \quad \text{for } j \neq \ell,\ 1 \le j,\ \ell \le N, \\ \text{(ii)} & x_j - y_j = d \in \mathbb{R}^s\backslash\{0\} \quad \text{for } 1 \le j \le N. \end{array}$$

Definition 3.1. Let F: $\mathbb{R}^s \to \mathbb{R}$, $F \in RP_0^\infty$ or RN_1^∞, and set $A(F)$ to be the $N \times N$ matrix given by

$$(3.2)\qquad A(F) := \{F(y_j - x_\ell)\}_{j,\ell=1}^N.$$

Furthermore, let

$$(3.3)\qquad \begin{aligned} A_1(F) &:= \frac{A(F)+A(F)^T}{2}, \\ A_2(F) &:= \frac{A(F)-A(F)^T}{2}, \end{aligned}$$

and

$$U_N := \left\{ \alpha = (\alpha_1, \alpha_2, \dots, \alpha_N) \in \mathbb{R}^N : \quad \sum_{j=1}^N \alpha_j = 0 \right\}.$$

The primary focus of interest in this section is the quadratic form $\langle A_1(F)\alpha, \alpha\rangle$, $\alpha \in \mathbb{R}^N$, associated with F in RP_0^∞ and RN_1^∞. (Throughout this paper, our matrices will act on real vectors only.) We denote this quadratic form by $Q_0(F)$ if $F \in RP_0^\infty$, and by $Q_1(F)$ if $F \in RN_1^\infty$. The analysis of $Q_0(F)$ and $Q_1(F)$—which will supply us with a useful tool—begins with the following.

Theorem 3.2. *Let $F \in RP_0^\infty$ or RN_1^∞, and let $d\mu$ be its representing measure. Suppose that $\alpha = (\alpha_1, \dots, \alpha_N) \in \mathbb{R}^N$ and that $Q_0(F)$ and $Q_1(F)$ are defined as above. Then for $m = 0, 1$, the following holds:*

$$\begin{aligned} Q_m(F) &= \sum_{j,\ell=1}^N \alpha_j\alpha_\ell F(x_j - x_\ell) \\ &- \frac{2(-1)^m}{(2\pi)^s}\int_0^\infty \frac{1}{t^{s/2+m}}\int_{\mathbb{R}^s} e^{-\|u\|^2/4t}\sin^2\left(\frac{u\cdot d}{2}\right)\left|\sum_{j=1}^N \alpha_j e^{ix_j\cdot u}\right|^2 du\,d\mu(t) \\ &=: I(\mu) - (-1)^m J(\mu, d). \end{aligned}$$

Proof. Let $m = 0$ or 1. From Definition 3.1,

$$
\begin{aligned}
Q_m(F) &= \sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell \left[\frac{F(y_j - x_\ell) + F(y_\ell - x_j)}{2} \right] \\
&= \sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell F(x_j - x_\ell) - \sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell \\
&\quad \times \left[\frac{F(x_j - x_\ell) - F(y_j - x_\ell) + F(x_\ell - x_j) - F(y_\ell - x_j)}{2} \right] \\
&= \sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell F(x_j - x_\ell) - (-1)^m \sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell \int_0^\infty \frac{1}{t^m} \\
&\quad \times \left[\frac{e^{-\|x_j - x_\ell\|^2 t} - e^{-\|y_j - x_\ell\|^2 t} + e^{-\|x_\ell - x_j\|^2 t} - e^{-\|y_\ell - x_j\|^2 t}}{2} \right] d\mu(t),
\end{aligned}
\tag{3.4}
$$

where the last step follows from (2.1) and (2.2). Using now the fact that

$$
e^{-\|\xi\|^2 t} = \frac{1}{(2\pi)^s t^{s/2}} \int_{\mathbb{R}^s} e^{-\|u\|^2/4t} e^{-i\xi\cdot u} du, \qquad t > 0, \quad \xi \in \mathbb{R}^s,
$$

(3.1) and (3.4), we see that

$$
\begin{aligned}
Q_m(F) &= \sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell F(x_j - x_\ell) - \frac{(-1)^m}{(2\pi)^s} \sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell \int_0^\infty \frac{1}{t^{s/2+m}} \int_{\mathbb{R}^s} e^{-\|u\|^2/4t} \\
&\quad \times \left[\frac{e^{-i(x_j - x_\ell)\cdot u} - e^{-i(x_j - x_\ell - d)\cdot u} + e^{-i(x_\ell - x_j)\cdot u} - e^{-i(x_\ell - x_j - d)\cdot u}}{2} \right] du d\mu(t) \\
&= \sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell F(x_j - x_\ell) - \frac{(-1)^m}{(2\pi)^s} \int_0^\infty \frac{1}{t^{s/2+m}} \int_{\mathbb{R}^s} e^{-\|u\|^2/4t} \\
&\quad \times \left[\sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell \left[\frac{e^{-i(x_j - x_\ell)\cdot u}\{1 - e^{id\cdot u}\} + e^{-i(x_\ell - x_j)\cdot u}\{1 - e^{id\cdot u}\}}{2} \right] \right] du d\mu(t) \\
&= \sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell F(x_j - x_\ell) \\
&\quad - \frac{(-1)^m}{(2\pi)^s} \int_0^\infty \frac{1}{t^{s/2+m}} \int_{\mathbb{R}^s} e^{-\|u\|^2/4t} \\
&\quad \times \left[\sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell e^{-i(x_j - x_\ell)\cdot u} \{1 - e^{id\cdot u}\} \right] du d\mu(t).
\end{aligned}
\tag{3.5}
$$

Since the numbers $Q_m(F)$ and

$$
\sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell e^{-i(x_j - x_\ell)\cdot u} = \left| \sum_{j=1}^{N} \alpha_j e^{i x_j \cdot u} \right|^2 = \sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell \cos[u \cdot (x_j - x_\ell)]
$$

are real, it follows that

$$\begin{aligned} Re &\left[\sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell e^{-i(x_j-x_\ell)\cdot u}\{1 - e^{idu}\}\right] \\ &= [1-\cos(d\cdot u)]\left[\sum_{j,\ell=1}^{N} \alpha_j\alpha_\ell e^{-i(x_j-x_\ell)\cdot u}\right] \\ &= 2\sin^2\left(\frac{u\cdot d}{2}\right)\left|\sum_{j=1}^{N}\alpha_j e^{ix_j\cdot u}\right|^2 . \end{aligned}$$

Consequently, we conclude from (3.5) that

$$\begin{aligned} Q_m(F) &= \sum_{j,\ell=1}^{N} \alpha_j\alpha_\ell F(x_j - x_\ell) \\ &- \frac{2(-1)^m}{(2\pi)^s}\int_0^\infty \frac{1}{t^{s/2+m}}\int_{\mathbb{R}^s} e^{-\|u\|^2/4t}\sin^2\left(\frac{u\cdot d}{2}\right)\left|\sum_{j=1}^{N}\alpha_j e^{ix_j\cdot u}\right|^2 du d\mu(t), \end{aligned}$$

as claimed. □

Remark 3.3. The expression $I(\mu)$ given in Theorem 3.2 has been studied in detail in [NW2]. In particular, the following theorem was established there.

THEOREM 3.4. *Let $F \in RP_0^\infty$ or RN_1^∞, and let $d\mu$ be its representing measure. Let $\Gamma(z)$ denote the standard gamma function, $2q = \min_{j\neq\ell}\|x_j - x_\ell\|$,*

$$\delta := 12\left[\frac{\pi\Gamma^2\left(\frac{s+2}{2}\right)}{9}\right]^{1/(s+1)} \quad \text{and} \quad C_s := \frac{\delta^s}{2^{s+1}\Gamma((s+2)/2)}.$$

Then

$$\frac{(-1)^m I(\mu)}{\|\alpha\|^2} \geq \frac{C_s}{q^s}\int_0^\infty \frac{e^{-\delta^2/q^2 t}}{t^{s/2+m}} d\mu(t) =: I'(\mu),$$

where $\alpha = (\alpha_1, \ldots, \alpha_N) \in \mathbb{R}^N\backslash\{0\}$ is arbitrary if $F \in RP_0^\infty (m = 0)$ and belongs to U_N if $F \in RN_1^\infty (m = 1)$.

Remark 3.5. (i) Suppose that $J(\mu, d)$ is given by Theorem 3.2. Since

$$\left|\sum_{j=1}^{N}\alpha_j e^{iu\cdot x_j}\right|^2 = \sum_{j,\ell=1}^{N}\alpha_j\alpha_\ell\cos[u\cdot(x_j - x_\ell)] \leq N\|\alpha\|^2, \qquad \alpha \in \mathbb{R}^N,$$

we see that

$$\begin{aligned} 0 \leq J(\mu, d) &\leq N\|\alpha\|^2\left[\frac{2}{(2\pi)^s}\int_{\mathbb{R}^s}\sin^2\left(\frac{u\cdot d}{2}\right)\left[\int_0^\infty \frac{e^{-\|u\|^2/4t}}{t^{s/2+m}} d\mu(t)\right] du\right] \\ &=: N\|\alpha\|^2\tilde{J}(\mu, d). \end{aligned}$$

(ii) In certain specific cases, it is possible to estimate $J(\mu, d)/\|\alpha\|^2$ from above by a quantity independent of N (see Proposition 3.11 and also §5).

THEOREM 3.6. *Let $F \in RP_0^\infty$ or RN_1^∞, and let $d\mu$ be its representing measure. Suppose that $F(0) \geq 0$ if $F \in RN_1^\infty$ and that $I'(\mu) > N\widetilde{J}(\mu, d)$. Then*

$$\|A_1(F)\alpha\| \geq [I'(\mu) - N\widetilde{J}(\mu, d)]\|\alpha\|, \qquad \alpha \in \mathbb{R}^N.$$

Proof. If $F \in RP_0^\infty$, then on the one hand, Theorems 3.2 and 3.4 and Remark 3.5 show that

$$\langle A_1(F)\alpha, \alpha\rangle \geq [I'(\mu) - N\widetilde{J}(\mu, d)]\|\alpha\|^2, \qquad \alpha \in \mathbb{R}^N.$$

On the other hand, $\langle A_1(F)\alpha, \alpha\rangle \leq \|A_1(F)\alpha\|\|\alpha\|$ by the Cauchy–Schwarz inequality. The desired result is now immediate.

Suppose now that $F \in RN_1^\infty$ and $\alpha \in U_N$. Using Theorems 3.2 and 3.4 and Remark 3.5 once again, we deduce that

$$\begin{aligned} -\langle A_1(F)\alpha, \alpha\rangle &= -I(\mu) - J(\mu, d) \\ &\geq [I'(\mu) - N\widetilde{J}(\mu, d)]\|\alpha\|^2, \qquad \alpha \in U_N. \end{aligned}$$

This guarantees that $A_1(F)$ has $N-1$ negative eigenvalues while the condition $F(0) \geq 0$ and (2.2) ensure that the trace of $A_1(F)$ is nonnegative. Thus $A_1(F)$ is conditionally negative definite and the required result now follows from [B]. □

We now turn to some specific illustrations involving the functions F_i, $1 \leq i \leq 4$.

PROPOSITION 3.7. *Let F be one of the functions F_i, $i = 1, 2, 3, 4$, and denote the corresponding expressions $I(\mu)$ by I_i, $1 \leq i \leq 4$, respectively. Suppose that $\alpha = (\alpha_1, \ldots, \alpha_N) \in R^N\backslash\{0\}$ is subject to the restriction that $\alpha \in U_N$ if $i = 1, 2$, but is arbitrary if $i = 3, 4$. Then the following estimates hold:*

(i) $-I_1/\|\alpha\|^2 \geq (C_s\Gamma((s+1)/2)/2\delta^{s+1}\sqrt{\pi})q =: I_1'$;

(ii) $-I_2/\|\alpha\|^2 \geq (C_s/2\delta^{(s+3)/2})(e^{-2\delta/q}/q^{(s-3)/2}\sqrt{\pi}) =: I_2'$;

(iii) *If* $q < 1$, *then* $I_3/\|\alpha\|^2 \geq (C_s/\delta^{(s+1)/2}\sqrt{\pi})a^{(s-3)/2}[1-e^{-a\delta}]\, e^{-\frac{\delta}{q}(1/a+a)}/q^{(s-1)/2}$ $=: I_3'$, *where a is an absolute constant for which $\delta > 2a$;*

(iv) $I_4/\|\alpha\|^2 \geq C_s e^{-\delta^2/(q^2\rho)}/q^s\rho^{s/2} =: I_4'$.

Proof. (i) Recall that $d\mu_1(t) = \frac{1}{2}(\pi t)^{-1/2}dt$. Using this in Theorem 3.4, we obtain

$$\begin{aligned} \frac{-I_1}{\|\alpha\|^2} &\geq \frac{C_s}{2q^s\sqrt{\pi}} \int_0^\infty \frac{e^{-\delta^2/q^2t}}{t^{(s+3)/2}}dt \\ &= C_s\frac{q}{2\sqrt{\pi}} \int_0^\infty e^{-\delta^2 u}u^{(s-1)/2}du \quad \left(\text{where } u = \frac{1}{q^2t}\right) \\ &= \frac{C_s\Gamma((s+1)/2)\,q}{2\delta^{s+1}\sqrt{\pi}}. \end{aligned}$$

(ii) By Theorem 3.4, and the fact that $d\mu_2(t) = \frac{1}{2}(\pi t)^{-1/2}e^{-t}dt$,

$$\frac{-I_2}{\|\alpha\|^2} \geq \frac{C_s}{2q^s\sqrt{\pi}} \int_0^\infty \frac{e^{-\delta^2/q^2t}e^{-t}}{t^{(s+3)/2}}dt$$

$$= \frac{C_s q}{2\sqrt{\pi}} \int\limits_0^\infty e^{-\delta^2 u} e^{-1/q^2 u} u^{(s-1)/2} du \qquad \left(\text{by setting } u = \frac{1}{q^2 t}\right)$$

$$\geq \frac{C_s q}{2\sqrt{\pi}} \int\limits_{1/\delta q}^\infty e^{-\delta^2 u} e^{-1/q^2 u} u^{(s-1)/2} du$$

$$\geq \frac{C_s q e^{-\delta/q}}{2(\delta q)^{(s-1)/2}\sqrt{\pi}} \int\limits_{1/\delta q}^\infty e^{-\delta^2 u} du$$

(as $u \mapsto u^{(s-1)/2}$ and $u \mapsto e^{-1/q^2 u}$ are increasing functions)

$$= \frac{C_s}{2\delta^{(s+3)/2}\sqrt{\pi}} \frac{e^{-2\delta/q}}{q^{(s-3)/2}}.$$

(iii) By Stirling's formula, $\lim_{s\to\infty} \delta/(s+2) = 6/e$, so there exists a positive absolute constant a such that $\delta > 2a$. Now, since $d\mu_3(t) = (\pi t)^{-1/2} e^{-t} dt$ and $m = 0$, we get

$$\frac{I_3}{\|\alpha\|^2} \geq \frac{C_s}{\sqrt{\pi} q^s} \int\limits_0^\infty e^{-\delta^2/q^2 t} \frac{e^{-t}}{t^{(s+1)/2}} dt$$

$$= \frac{C_s}{\sqrt{\pi} q} \int\limits_0^\infty e^{-\delta^2 u} \frac{e^{-1/q^2 u}}{u} u^{(s-1)/2} du \qquad \left(u = \frac{1}{q^2 t}\right).$$

As the function $u \mapsto (e^{-1/q^2 u})/u$ increases on $[0, 1/q^2]$, and $q < 1$, we see that

$$\frac{I_3}{\|\alpha\|^2} \geq \frac{C_s}{q\sqrt{\pi}} \int\limits_{a/\delta q}^{2a/\delta q} e^{-\delta^2 u} \frac{e^{-1/q^2 u}}{u} u^{(s-1)/2} du$$

$$\geq \frac{C_s}{q\sqrt{\pi}} \frac{\delta q}{a} e^{-\delta/(aq)} \left(\frac{a}{\delta q}\right)^{(s-1)/2} \int\limits_{a/\delta q}^{2a/\delta q} e^{-\delta^2 u} du$$

$$\geq \frac{C_s}{\delta^{(s+1)/2}} \frac{a^{(s-3)/2}}{\sqrt{\pi}} (1 - e^{-\frac{a\delta}{q}}) \frac{e^{-\delta/q((1/a)+a)}}{q^{(s-1)/2}}$$

$$\geq \frac{C_s}{\delta^{(s+1)/2}} \frac{a^{(s-3)/2}}{\sqrt{\pi}} (1 - e^{-a\delta}) \frac{e^{-\delta/q((1/a)+a)}}{q^{(s-1)/2}}.$$

(iv) Recall that $m = 0$ and $d\mu_4(t) = \delta(t - \rho)dt$. So

$$\frac{C_s}{q^s} \int_0^\infty \frac{e^{-\delta^2/q^2 t}}{t^{s/2+m}} d\mu_4(t) = \frac{C_s e^{-\delta^2/(q^2 \rho)}}{q^s \rho^{s/2}}. \qquad \square$$

PROPOSITION 3.8. *Let $\widetilde{F}$ be one of the functions F_i, $1 \leq i \leq 3$, and let the corresponding expressions $\widetilde{J}(\mu, d)$ be denoted by $\widetilde{J}_i$, $1 \leq i \leq 3$. Then the following hold:*

(i) $\widetilde{J}_2, \widetilde{J}_3 \leq \widetilde{J}_1$;

(ii) $\widetilde{J}_1 \leq (2\Gamma((s+1)/2)\,\omega_{s-1}/\pi^{s+1/2})\|d\| =: J'\|d\|$,

where ω_{s-1} denotes the $s-1$-dimensional volume of the unit sphere in $\mathbb{R}^s$.

Proof. (i) $\widetilde{J}_2 \le \widetilde{J}_1$ because $d\mu_2(t) \le d\mu_1(t)$.

Next, observe that

$$\begin{aligned}\widetilde{J}_3 &= \frac{2}{(2\pi)^s}\int_{\mathbb{R}^s} \sin^2\left(\frac{u\cdot d}{2}\right)\left[\frac{1}{\sqrt{\pi}}\int_0^\infty e^{-\|u\|^2/4t}\frac{e^{-t}}{t^{(s+1)/2}}dt\right] du\\ &= \frac{2}{(2\pi)^s}\int_{\mathbb{R}^s} \sin^2\left(\frac{u\cdot d}{2}\right)\left[\frac{1}{\sqrt{\pi}}\int_0^\infty (te^{-t})\frac{e^{-\|u\|^2/4t}}{t^{(s+3)/2}}dt\right] du\\ &\le 2e^{-1}\frac{2}{(2\pi)^s}\int_{\mathbb{R}^s} \sin^2\left(\frac{u\cdot d}{2}\right)\left[\frac{1}{2\sqrt{\pi}}\int_0^\infty \frac{e^{-\|u\|^2/4t}}{t^{(s+3)/2}}dt\right] du\\ &= 2e^{-1}\widetilde{J}_1 < \widetilde{J}_1.\end{aligned}$$

(ii) Employing the substitution $w = 1/4t$ in the expression for $\widetilde{J}_1$ above leads to

$$\begin{aligned}\widetilde{J}_1 &= \frac{2^{s+1}}{(2\pi)^s\sqrt{\pi}}\int_{\mathbb{R}^s} \sin^2\left(\frac{u\cdot d}{2}\right)\left(\int_0^\infty e^{-\|u\|^2 w}w^{(s-1)/2}dw\right) du\\ &= \frac{2}{\pi^{s+1/2}}\Gamma\left(\frac{s+1}{2}\right)\int_{\mathbb{R}^s} \sin^2\left(\frac{u\cdot d}{2}\right)\frac{du}{\|u\|^{s+1}}\\ &\le \frac{2\Gamma\left((s+1)/2\right)}{\pi^{s+1/2}}\left[\frac{1}{4}\int_{\|u\|\le\frac{2}{\|d\|}}\frac{|u\cdot d|^2}{\|u\|^{s+1}}du + \int_{\|u\|>\frac{2}{\|d\|}}\frac{du}{\|u\|^{s+1}}\right]\\ &\le \frac{2\Gamma\left((s+1)/2\right)}{\pi^{s+1/2}}\left[\frac{1}{4}\|d\|^2\int_{\|u\|\le\frac{2}{\|d\|}}\frac{du}{\|u\|^{s-1}} + \int_{\|u\|>\frac{2}{\|d\|}}\frac{du}{\|u\|^{s+1}}\right],\end{aligned}$$

by the Cauchy–Schwarz inequality. Computing each of the two integrals above using polar coordinates, we see that

$$\begin{aligned}\widetilde{J}_1 &\le \frac{2\Gamma\left((s+1)/2\right)}{\pi^{s+1/2}}\left[\frac{1}{4}\|d\|^2\omega_{s-1}\int_0^{2/\|d\|}dr + \omega_{s-1}\int_{2/\|d\|}^\infty\frac{dr}{r^2}\right]\\ &= \frac{2\Gamma\left((s+1)/2\right)}{\pi^{s+1/2}}\omega_{s-1}\left[\frac{\|d\|}{2}+\frac{\|d\|}{2}\right] = \frac{2\Gamma\left((s+1)/2\right)\omega_{s-1}}{\pi^{s+1/2}}\|d\|. \qquad \square\end{aligned}$$

Theorem 3.6, taken in conjunction with Propositions 3.7 and 3.8, yields the following.

COROLLARY 3.9. *Suppose that I_i', $1 \le i \le 3$, and J' are given by Propositions* 3.7 *and* 3.8, *respectively. If $NJ'\|d\| < I_i'$, $1 \le i \le 3$, then*

$$\|A_1(F_i)\alpha\| \ge [I_i' - NJ'\|d\|]\|\alpha\|, \qquad 1 \le i \le 3, \quad \alpha \in \mathbb{R}^N.$$

Remark 3.10. (i) The function $F(x) := \|x\|^{2\beta}$, $x \in \mathbb{R}^s$, $0 < \beta < 1$, belongs to RN_1^∞ and is represented by the measure $d\mu(t) = \beta[\Gamma(1-\beta)]^{-1}t^{-\beta}\,dt$. We can show that the corresponding expressions $\widetilde{I}(\mu)$ and $\widetilde{J}(\mu, d)$ satisfy the following estimates:

$$\widetilde{I}(\mu) \ge \frac{C_s\beta\Gamma\left(\beta+(s/2)\right)}{[\Gamma(1-\beta)]\delta^{2\beta+s}}q^{2\beta} \quad \text{and} \quad \widetilde{J}(\mu,d) \le \frac{\Gamma\left((s+2\beta)/2\right)\omega_{s-1}\|d\|^{2\beta}}{\pi(1-\beta)\Gamma(1-\beta)}.$$

It is evident that the special case $\beta = \frac{1}{2}$ corresponds exactly to the function F_1.

(ii) Let $\widetilde{J}_2$ be given by Proposition 3.8. With a little more effort, we can show that

$$\widetilde{J}_2 \le \begin{cases} \dfrac{\omega_{s-1}}{\pi^s}\left[4\Gamma\left(\dfrac{s+1}{2}\right) + \displaystyle\sum_{j=0}^{p} \frac{p!}{j!}\Gamma(j+1)\right] \|d\|^2 & \text{if } s = 2p+1; \\ \dfrac{\omega_{s-1}}{\pi^s}\left[4\Gamma\left(\dfrac{s+1}{2}\right) + \displaystyle\sum_{j=0}^{p} \frac{p!}{j!}\Gamma\left(j+\frac{1}{2}\right)\right] \|d\|^2 & \text{if } s = 2p. \end{cases}$$

For future reference, we abbreviate this estimate by $\widetilde{J}_2 \le J''\|d\|^2$.

The next result deals with the Gaussian F_4; it pursues a somewhat different tack leading to a lower estimate for $\|A_1(F_4)\alpha\|$ that does not involve N.

PROPOSITION 3.11. *Suppose that I_4' is given by Proposition* 3.7 *and that $\alpha \in \mathbb{R}^N$. If $\|d\| \le q$, then*

$$\|A_1(F_4)\alpha\| \ge [I_4' - D_s(\rho, q)\|d\|]\, \|\alpha\|,$$

where

$$D_s(\rho, q) := \pi^{-s}\rho q \left[1 + 6s \sum_{n=0}^{\infty} (n+3)^s e^{-\rho n^2 q^2}\right].$$

Proof. As before, in view of the Cauchy–Schwarz inequality, it suffices to prove that

$$\langle A_1(F_4)\alpha, \alpha\rangle \ge [I_4' - D_s(\rho, q)\|d\|]\, \|\alpha\|^2.$$

Since $m = 0$ and $d\mu_4(t) = \delta(t-\rho)dt$, Theorem 3.2 implies

$$\begin{aligned} \langle A_1(F_4)\alpha, \alpha\rangle &= I_4 - \frac{2}{(2\pi)^s \rho^{s/2}} \int_{\mathbb{R}^s} e^{-\|u\|^2/4\rho} \sin^2\left(\frac{u\cdot d}{2}\right) \left|\sum_{j=1}^{N} \alpha_j e^{ix_j\cdot u}\right|^2 du \\ &=: I_4 - J_4. \end{aligned} \tag{3.6}$$

By Proposition 3.7 (iv),

$$I_4 \ge I_4'\|\alpha\|^2. \tag{3.7}$$

So it remains to show that

$$|J_4| \le D_s(\rho, q)\|d\|\, \|\alpha\|^2. \tag{3.8}$$

To this end, we write

$$\begin{aligned} J_4 &= \frac{2}{(2\pi)^s \rho^{s/2}} \int_{\mathbb{R}^s} e^{-\|u\|^2/4\rho} \sin^2\left(\frac{u\cdot d}{2}\right) \left[\sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell e^{-i(x_j - x_\ell)\cdot u}\right] du \\ &= 2\sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell \frac{1}{(2\pi)^s \rho^{s/2}} \int_{\mathbb{R}^s} e^{-\|u\|^2/4\rho} \sin^2\left(\frac{u\cdot d}{2}\right) e^{-i(x_j - x_\ell)\cdot u} du \qquad (3.9) \\ &= \frac{2}{\pi^s} \sum_{j,\ell=1}^{N} \alpha_j \alpha_\ell \frac{1}{(2\pi)^s} \int_{\mathbb{R}^s} \widehat{F_4}(u) \sin^2\left(\frac{u\cdot d}{2}\right) e^{i(x_\ell - x_j)\cdot u} du. \end{aligned}$$

Let

$$\Delta_d F_4(x) := F_4(x+d) - 2F_4(x) + F_4(x-d).$$

It is quite straightforward to verify that

$$(\widehat{\Delta_d F_4})(u) = -4\widehat{F_4}(u)\sin^2\left(\frac{u\cdot d}{2}\right);$$

so from (3.9),

$$\begin{aligned} J_4 &= -\frac{1}{2\pi^s}\sum_{j,\ell=1}^{N}\alpha_j\alpha_\ell\frac{1}{(2\pi)^s}\int_{\mathbb{R}^s}(\widehat{\Delta_d F_4})(u)e^{i(x_\ell-x_j)\cdot u}du \\ &= -\frac{1}{2\pi^s}\sum_{j,\ell=1}^{N}\alpha_j\alpha_\ell(\Delta_d F_4)(x_\ell-x_j), \\ &= -\frac{1}{2\pi^s}\left[\sum_{j=1}^{N}|\alpha_j|^2(\Delta_d F_4)(0) + \sum_{\substack{j,\ell=1\\ j\neq\ell}}^{N}\alpha_j\alpha_\ell(\Delta_d F_4)(x_\ell-x_j)\right], \end{aligned}$$

whence

$$\begin{aligned} |J_4| &\le \frac{1}{2\pi^s}\left[\sum_{j=1}^{N}|\alpha_j|^2|(\Delta_d F_4)(0)| + \sum_{j,\ell=1}^{N}{}'|\alpha_j\alpha_\ell||(\Delta_d F_4)(x_\ell-x_j)|\right] \\ &=: \frac{1}{2\pi^s}[S_1+S_2], \end{aligned} \tag{3.10}$$

where $\sum'$ in the penultimate line indicates that $j\neq\ell$ in that sum. In order to estimate S_1, note that

$$|(\Delta_d F_4)(0)| = 2|1-e^{-\rho\|d\|^2}| \le 2\rho\|d\|^2,$$

so

$$S_1 \le 2\rho\|d\|^2\|\alpha\|^2. \tag{3.11}$$

We now take up S_2: observe that

$$\begin{aligned} S_2 &\le \sum_{j,\ell=1}^{N}{}'\frac{|\alpha_j|^2+|\alpha_\ell|^2}{2}|\Delta_d F_4(x_\ell-x_j)| \\ &\le \max\{\gamma_N,\tilde{\gamma}_N\}\|\alpha\|^2, \end{aligned} \tag{3.12}$$

where

$$\gamma_N := \max_j\sum_{\ell=1}^{N}{}'|(\Delta_d F_4)(x_\ell-x_j)| \quad\text{and}\quad \tilde{\gamma}_N := \max_\ell\sum_{j=1}^{N}{}'|(\Delta_d F_4)(x_\ell-x_j)|.$$

Let j_0 and ℓ_0 be the indices where the maximum values of γ_N and $\tilde{\gamma}_N$ are attained, respectively, and suppose that $y\in\mathbb{R}^s$ with $\|y\|>\|d\|$. Plainly,

$$\begin{aligned}|(\Delta_d F_4)(y)| &= \left|e^{-\rho\|y\|^2} - e^{-\rho\|y-d\|^2} + e^{-\rho\|y\|^2} - e^{-\rho\|y+d\|^2}\right| \\ &\le \left|e^{-\rho\|y\|^2} - e^{-\rho\|y-d\|^2}\right| + \left|e^{-\rho\|y\|^2} - e^{-\rho\|y+d\|^2}\right|.\end{aligned}$$

By the Mean Value Theorem,

$$\begin{aligned}\left|e^{-\rho\|y\|^2} - e^{-\rho\|y-d\|^2}\right| &\le \big|\,\|y\| - \|y-d\|\,\big|\,2\rho\max\{\|y\|, \|y-d\|\}e^{-\rho[\min\{\|y\|,\|y-d\|\}]^2} \\ &\le \|d\|2\rho[\|y\| + \|d\|]e^{-\rho[\|y\|-\|d\|]^2}\end{aligned}$$

and

$$\left|e^{-\rho\|y\|^2} - e^{-\rho\|y+d\|^2}\right| \le \|d\|2\rho[\|y\| + \|d\|]e^{-\rho[\|y\|-\|d\|]^2}.$$

Therefore,

$$\gamma_N = \sum_{\substack{\ell=1\\ \ell\neq j_0}}^{N} |(\Delta_d F_4)(x_\ell - x_{j_0})| \le 4\rho\|d\| \sum_{\substack{\ell=1\\ \ell\neq j_0}}^{N} [\|x_\ell - x_{j_0}\| + \|d\|]e^{-\rho[\|x_\ell - x_{j_0}\|-\|d\|]^2}.$$

Adapting at this stage an argument from [NW1, pp. 79–80] and noting that $\|x_\ell - x_{j_0}\| \ge 2q > \|d\|$, we conclude from the above that

$$\gamma_N \le 12s\rho\|d\| \sum_{n=1}^{\infty}(n+2)^{s-1}\kappa_n,$$

where

$$\begin{aligned}\kappa_n &:= \sup\{[\|y\| + \|d\|]e^{-\rho[\|y\|-\|d\|]^2}; nq \le \|y\| \le (n+1)q\} \\ &\le [(n+1)q + \|d\|]e^{-\rho[nq-\|d\|]^2} \le (n+2)qe^{-\rho(n-1)^2q^2}.\end{aligned}$$

Consequently,

$$\begin{aligned}\gamma_N &\le 12s\rho\|d\|q\sum_{n=1}^{\infty}(n+2)^s e^{-\rho(n-1)^2q^2} \\ &= 12s\rho\|d\|q\sum_{n=0}^{\infty}(n+3)^s e^{-\rho n^2q^2}.\end{aligned} \tag{3.13}$$

The same argument also shows

$$\tilde{\gamma}_N \le 12s\rho\|d\|q\sum_{n=0}^{\infty}(n+3)^s e^{-\rho n^2q^2}. \tag{3.14}$$

Thus, from (3.12), (3.13), and (3.14), it follows that

$$S_2 \le 12s\rho q\|d\|\ \|\alpha\|^2\sum_{n=0}^{\infty}(n+3)^s e^{-\rho n^2q^2}. \tag{3.15}$$

Finally, from (3.10), (3.11), and (3.15), we may conclude that

$$\begin{aligned}|J_4| &\leq \frac{1}{2\pi^s}\left[2\rho q\|d\| + 12s\rho q\|d\|\sum_{n=0}^{\infty}(n+3)^s e^{-\rho n^2 q^2}\right]\|\alpha\|^2\\ &\leq D_s(\rho, q)\|d\|\ \|\alpha\|^2,\end{aligned}$$

thereby completing the proof. □

It is natural to ask whether the procedure adopted in the preceding proposition can be adapted to other comparable situations. It turns out that while the method itself can be extended to certain other situations, it may not always yield better results. To convey a general idea of the matter, let us suppose that $\alpha \in U_N$, $F \in RN_1^\infty$, and that its representing measure is $d\mu$. Then, under fairly mild conditions on F—conditions that do not preclude any of the commonly used functions—it is shown in [Ba] that the generalized Fourier transform $\widehat{F}$ of F is given by

$$\widehat{F}(u) = -K_s\int_0^\infty e^{-\|u\|^2/4t}\frac{d\mu(t)}{t^{s/2+1}}, \qquad u \in \mathbb{R}^s\backslash\{0\}.$$

(Hereafter, K_s will denote a generic constant that depends on s but whose actual numerical value is likely to change from one appearance to another.) Consequently,

$$\begin{aligned}J(\mu, d) &= -K_s\int_{\mathbb{R}^s}\sin^2\left(\frac{u\cdot d}{2}\right)\left|\sum_{j=1}^{N}\alpha_j e^{iu\cdot x_j}\right|^2\widehat{F}(u)du\\ &= K_s\int_{\mathbb{R}^s}(\Delta_d(F(\cdot)))^\wedge(u)\left|\sum_{j=1}^{N}\alpha_j e^{iu\cdot x_j}\right|^2 du,\end{aligned}$$

where, as before, $\Delta_d F(x) = F(x+d) + F(x-d) - 2F(x)$. Again, from [Ba] (noting that if F satisfies the mild conditions mentioned above, then so does $\Delta_d F$), we see that

$$\begin{aligned}J(\mu, d) &= K_s\sum_{j,\ell=1}^{N}\alpha_j\alpha_\ell\Delta_d F(x_j - x_\ell)\\ &= K_s\|\alpha\|^2[F(d) - F(0)] + K_s\sum_{j\neq\ell}\alpha_j\alpha_\ell\Delta_d F(x_j - x_\ell).\end{aligned}$$

The packing argument from [NW1, p. 79] (as employed in Proposition 3.11) once again leads to

$$|J(\mu, d)| \leq K_s\|\alpha\|^2\left[|F(d) - F(0)| + \sum_{n=1}^{\infty}n^{s-1}\kappa_n(d)\right],$$

where $\kappa_n(d) := \sup\{|\Delta_d F(x)|\colon nq \leq \|x\| < (n+1)q\}$. Now, the point is that if we expect to get a good estimate for $J(\mu, d)$ in terms of a quantity that is independent of N and which approaches zero along with $\|d\|$, then $\kappa_n(d)$ should be very well behaved. Indeed, this does happen for the Gaussian and possibly other functions which decay sufficiently rapidly at infinity. However, $\kappa_n(d)$ may not always exhibit such desirable behaviour; in fact, even for $s = 2$ and $F = F_1$, the rate of decay of $\kappa_n(d)$ is not rapid enough to suit our purposes.

4. Main results. Let $X = \{x_1, \ldots, x_N\}$ be a set of distinct centers and $2q = \min_{j \neq \ell} \|x_j - x_\ell\|$. For each center x_i, recall that $W(x_i)$ is the s-dimensional cube centered at x_i and with side length $2qs^{-1/2}$, i.e.,

$$W(x_i) := \{x \in \mathbb{R}^s : \ \|x - x_i\|_\infty \leq qs^{-1/2}\},$$

where $\| \ \|_\infty$ denotes the supremum norm in $\mathbb{R}^s$. Note that $\|x - x_i\| \leq q$ for all $x \in W(x_i)$ and, therefore, the cubes $W(x_i)$ are essentially disjoint. Recall also that for the given set X of centers, a closed and bounded domain $\Omega \subset \mathbb{R}^s$ is called admissible with respect to X, if $W(x_i) \subset \Omega$ for all $i = 1, \ldots, N$.

We are interested in the problem of finding the best L^2 approximation to a given $f \in C(\Omega)$, from a linear subspace of functions in $C(\Omega)$. The latter space will be the span of the functions $F(\cdot - x_i)$, where F is an appropriate radial function.

DEFINITION 4.1. Let $F \in RP_0^\infty$ or RN_1^∞, and suppose that Ω is an admissible domain with respect to the set $X = \{x_1, \ldots, x_N\}$ of distinct centers. The N-dimensional linear space $S_X \subset C(\Omega)$ is defined as

$$S_X := \ \text{span}\{F(x - x_i) : 1 \leq i \leq N\}. \tag{4.1}$$

DEFINITION 4.2. For a given function $f \in C(\Omega)$ and the space S_X as in Definition 4.1, the continuous least squares approximation problem is to find the best L_2 approximant to f from S_X, i.e., to find a function $s^* \in S_X$ such that

$$\|f - s^*\| = \inf_{s \in S_X} \|f - s\|.$$

Here $\|f\| = \langle f, f\rangle^{\frac{1}{2}} = \left(\int_\Omega f(x)^2 dx\right)^{\frac{1}{2}}$ is the usual L_2 norm and $\langle f, g\rangle$ is the corresponding inner product $\int_\Omega f(x)g(x)dx$.

As is well known, the best L_2 approximant to $f \in C(\Omega)$ can be found by solving the linear system of the so-called *normal equations*:

$$A_X \alpha = z, \tag{4.2}$$

where $\alpha = (\alpha_i)_{1 \leq i \leq N} \in \mathbb{R}^N$, and A_X is an $N \times N$ matrix with

$$A_X = \left(\langle F(\cdot - x_i), F(\cdot - x_j)\rangle\right)_{1 \leq i,j \leq N} \quad \text{and} \quad z = \left(\langle F(\cdot - x_i), f\rangle\right)_{1 \leq i \leq N}.$$

Our aim is to estimate $\langle A_X \alpha, \alpha\rangle$ from below. We begin with the following.

LEMMA 4.3.

$$\langle A_X \alpha, \alpha\rangle = \left\| \sum_{j=1}^N \alpha_j F(\cdot - x_j) \right\|^2 \geq \sum_{k=1}^N \int_{W(x_k)} \left| \sum_{j=1}^N \alpha_j F(x - x_j) \right|^2 dx.$$

Proof. From the definition of A_X, we obtain

$$\begin{aligned} \langle A_X \alpha, \alpha\rangle &= \sum_{i,j=1}^N \alpha_i \alpha_j \int_\Omega F(x - x_i) F(x - x_j)\, dx \\ &= \int_\Omega \left(\sum_{i=1}^N \alpha_i F(x - x_i)\right) \left(\sum_{j=1}^N \alpha_j F(x - x_j)\right) dx \end{aligned} \tag{4.3}$$

$$
\begin{aligned}
&= \int_\Omega \left| \sum_{j=1}^N \alpha_j F(x - x_j) \right|^2 dx \\
&\geq \int_{\bigcup_{k=1}^N W(x_k)} \left| \sum_{j=1}^N \alpha_j F(x - x_j) \right|^2 dx \\
&= \sum_{k=1}^N \int_{W(x_k)} \left| \sum_{j=1}^N \alpha_j F(x - x_j) \right|^2 dx. \qquad \square
\end{aligned}
$$

We want to express the integral over each cube $W(x_k)$ as the limit of certain Riemann sums whose point evaluations are taken over an equally spaced grid of points in $W(x_k)$. In view of Lemma 4.3, this will ultimately allow us to relate lower bounds on the quadratic form of the least squares matrix to those of certain interpolation matrices.

In what follows, let $\mathbf{i} = (i_1, \dots, i_s) \in \mathbb{Z}_+^s$ be an index vector, and define $\mathbf{e} = (1, 1, \dots, 1)$. By $\mathbf{i} \leq \mathbf{j}$, we mean $i_\ell \leq j_\ell$ for $\ell = 1, \dots, s$.

LEMMA 4.4. *Let X and Ω be as before. If $\alpha \in \mathbb{R}^N$, then*

$$
\langle A_X \alpha, \alpha \rangle \geq \underline{\lim}_{r \to \infty} \frac{2q^s s^{-s/2}}{r^s} \sum_{\substack{-r\mathbf{e} \leq \mathbf{i} \leq (r-1)\mathbf{e} \\ 0 \leq i_1 \leq r-1}} \|A_1^{(\mathbf{i})}(F)\alpha\|^2,
$$

where

$$
\begin{aligned}
A_1^{(\mathbf{i})}(F) &:= \frac{1}{2}(A^{(\mathbf{i})}(F) + A^{(\mathbf{i})}(F)^T), \\
A^{(\mathbf{i})}(F) &:= (F(y_k^{(\mathbf{i})} - x_j))_{1 \leq k,j \leq N}, \\
y_k^{(\mathbf{i})} &:= x_k + \frac{qs^{-1/2}}{r}\left(\mathbf{i} + \frac{1}{2}\mathbf{e}\right), \qquad -r\mathbf{e} \leq \mathbf{i} \leq (r-1)\mathbf{e}.
\end{aligned}
$$

Proof. Consider the following equally spaced grid of points:

$$
x_k + \frac{qs^{-1/2}}{r}\mathbf{i} \in W(x_k), \qquad -r\mathbf{e} \leq \mathbf{i} \leq r\mathbf{e},
$$

where $1 \leq k \leq N$ and $r \in \mathbb{N}_+^s$.

As the function F is continuous, the integral (4.3) can be recovered as the limit of Riemann sums. We choose as points of evaluation, the centers of the refined cubes in $W(x_k)$, namely,

$$
y_k^{(\mathbf{i})} := x_k + \frac{qs^{-1/2}}{r}\left(\mathbf{i} + \frac{1}{2}\mathbf{e}\right), \qquad -r\mathbf{e} \leq \mathbf{i} \leq (r-1)\mathbf{e}.
$$

As the side length of the refined cubes is $qs^{-1/2}r^{-1}$, we obtain

$$
\begin{aligned}
\langle A_X \alpha, \alpha \rangle &\geq \sum_{k=1}^N \underline{\lim}_{r \to \infty} \frac{q^s s^{-s/2}}{r^s} \sum_{-r\mathbf{e} \leq \mathbf{i} \leq (r-1)\mathbf{e}} \left| \sum_{j=1}^N \alpha_j F(y_k^{(\mathbf{i})} - x_j) \right|^2 \\
&= \underline{\lim}_{r \to \infty} \frac{q^s s^{-s/2}}{r^s} \sum_{-r\mathbf{e} \leq \mathbf{i} \leq (r-1)\mathbf{e}} \sum_{k=1}^N \left| \sum_{j=1}^N \alpha_j F(y_k^{(\mathbf{i})} - x_j) \right|^2 .
\end{aligned}
$$

A closer inspection of this representation reveals that it involves the matrices

$$A^{(\mathbf{i})}(F) = (F(y_k^{(\mathbf{i})} - x_j))_{1\le k,j\le N},$$

which are $N \times N$ *interpolation* matrices for radial interpolation at the points $y_k^{(\mathbf{i})}$. More precisely,

$$\begin{aligned}\langle A_X\alpha,\alpha\rangle &\ge \underline{\lim}_{r\to\infty}\frac{q^s s^{-s/2}}{r^s}\sum_{-r\mathbf{e}\le\mathbf{i}\le(r-1)\mathbf{e}}\sum_{k=1}^{N}\left|\sum_{j=1}^{N}\alpha_j F(y_k^{(\mathbf{i})}-x_j)\right|^2\\ &= \underline{\lim}_{r\to\infty}\frac{q^s s^{-s/2}}{r^s}\sum_{-r\mathbf{e}\le\mathbf{i}\le(r-1)\mathbf{e}}\|A^{(\mathbf{i})}(F)\alpha\|^2.\end{aligned}$$

The interpolation matrices which are (in general) not symmetric can be split up into a symmetric and a skew symmetric part, i.e.,

$$\begin{aligned}A^{(\mathbf{i})}(F) &= A_1^{(\mathbf{i})}(F) + A_2^{(\mathbf{i})}(F), \quad \text{where}\\ A_1^{(\mathbf{i})}(F) &= \frac{(A^{(\mathbf{i})}(F) + A^{(\mathbf{i})}(F)^T)}{2}, \quad \text{and}\\ A_2^{(\mathbf{i})}(F) &= \frac{(A^{(\mathbf{i})}(F) - A^{(\mathbf{i})}(F)^T)}{2}.\end{aligned}$$

Now the matrices $A_1^{(\mathbf{i})}(F)$ and $A_2^{(\mathbf{i})}(F)$ warrant further scrutiny. For any $\mathbf{i}$ with $-r\,\mathbf{e} \le \mathbf{i} \le (r-1)\,\mathbf{e}$, we have from the radiality of F,

$$\begin{aligned}&A_1^{(\mathbf{i})}(F)\\ &= \left(\frac{1}{2}\left(F(y_k^{(\mathbf{i})} - x_j) + F(y_j^{(\mathbf{i})} - x_k)\right)\right)_{1\le j,k\le N}\\ &= \left(\frac{1}{2}\left(F\left(x_k + \frac{qs^{-1/2}}{r}\left(\mathbf{i}+\frac{1}{2}\mathbf{e}\right) - x_j\right)\right.\right.\\ &\qquad\left.\left.+\ F\left(x_j + \frac{qs^{-1/2}}{r}\left(\mathbf{i}+\frac{1}{2}\mathbf{e}\right) - x_k\right)\right)\right)_{1\le j,k\le N}\\ &= \left(\frac{1}{2}\left(F\left(x_j + \frac{qs^{-1/2}}{r}\left(-\mathbf{i}-\frac{1}{2}\mathbf{e}\right) - x_k\right)\right.\right.\\ &\qquad\left.\left.+\ F\left(x_k + \frac{qs^{-1/2}}{r}\left(-\mathbf{i}-\frac{1}{2}\mathbf{e}\right) - x_j\right)\right)\right)_{1\le j,k\le N}\\ &= \left(\frac{1}{2}\left(F\left(x_k + \frac{qs^{-1/2}}{r}\left(-(\mathbf{i}+\mathbf{e})+\frac{1}{2}\mathbf{e}\right) - x_j\right)\right.\right.\\ &\qquad\left.\left.+\ F\left(x_j + \frac{qs^{-1/2}}{r}\left(-(\mathbf{i}+\mathbf{e})+\frac{1}{2}\mathbf{e}\right) - x_k\right)\right)\right)_{1\le j,k\le N}\\ &= \left(\frac{1}{2}\left(F(y_k^{(-\mathbf{i}-\mathbf{e})} - x_j) + F(y_j^{(-\mathbf{i}-\mathbf{e})} - x_k)\right)\right)_{1\le j,k\le N} = A_1^{(-\mathbf{i}-\mathbf{e})}(F).\end{aligned}$$

In a similar manner, it can be shown that

$$A_2^{(\mathbf{i})}(F) = -A_2^{(-\mathbf{i}-\mathbf{e})}(F).$$

Consequently,

$$\begin{aligned}\|A^{(\mathbf{i})}(F)\alpha\|^2+\|A^{(-\mathbf{i}-\mathbf{e})}(F)\alpha\|^2&=\|A_1^{(\mathbf{i})}(F)\alpha+A_2^{(\mathbf{i})}(F)\alpha\|^2+\|A_1^{(\mathbf{i})}(F)\alpha-A_2^{(\mathbf{i})}(F)\alpha\|^2\\&=2\|A_1^{(\mathbf{i})}(F)\alpha\|^2+2\|A_2^{(\mathbf{i})}(F)\alpha\|^2,\end{aligned}$$

where the last equality follows from the parallelogram identity.

Therefore, if Σ^* denotes the sum over the set of multi-indices $-r\mathbf{e}\le\mathbf{i}\le(r-1)\mathbf{e}$, $0\le i_1\le r-1$, and Σ^{**} denotes the sum taken over the set $-r\mathbf{e}\le\mathbf{i}\le(r-1)\mathbf{e}$, $-r\le i_1\le -1$, then

$$\begin{aligned}\sum_{-r\mathbf{e}\le\mathbf{i}\le(r-1)\mathbf{e}}\|A^{(\mathbf{i})}(F)\alpha\|^2&=\sum{}^{*}\|A^{(\mathbf{i})}(F)\alpha\|^2+\sum{}^{**}\|A^{(\mathbf{i})}(F)\alpha\|^2\\&=\sum{}^{*}\|A^{(\mathbf{i})}(F)\alpha\|^2+\sum{}^{*}\|A^{(-\mathbf{i}-\mathbf{e})}(F)\alpha\|^2\\&=\sum{}^{*}\left(\|A^{(\mathbf{i})}(F)\alpha\|^2+\|A^{(-\mathbf{i}-\mathbf{e})}(F)\alpha\|^2\right)\\&=\sum{}^{*}2\left(\|A_1^{(\mathbf{i})}(F)\alpha\|^2+\|A_2^{(\mathbf{i})}(F)\alpha\|^2\right)\\&\ge 2\sum{}^{*}\|A_1^{(\mathbf{i})}(F)\alpha\|^2,\end{aligned}$$

and finally,

$$\langle A_X\alpha,\alpha\rangle\ge\underline{\lim}_{r\to\infty}\frac{2q^s s^{-s/2}}{r^s}\sum_{\substack{-r\mathbf{e}\le\mathbf{i}\le(r-1)\mathbf{e}\\0\le i_1\le r-1}}\|A_1^{(\mathbf{i})}(F)\alpha\|^2.\qquad\square$$

We shall now investigate a subset of the grid points $y_k^{(\mathbf{i})}$ which are close to the points x_j. As a consequence, we shall obtain lower bounds on the quadratic form of the least squares matrix in terms of lower bounds on the quadratic form of certain interpolation matrices.

LEMMA 4.5. *Suppose* $0<\varepsilon\le 1$ *is fixed. Let*

$$P_\varepsilon(r):=\{\mathbf{i}:\|\mathbf{i}\|_\infty\le\varepsilon r-\tfrac12\text{ and }i_1\ge 0\}\quad\text{for }r\in\mathbb{N}.$$

Then the following statements hold:

(i) *There is an index* $r_0(\varepsilon)\in\mathbb{N}$ *such that* $P_\varepsilon(r)\neq\emptyset$ *for all* $r\ge r_0(\varepsilon)$;

(ii) $P_\varepsilon(r)\subset\{\mathbf{i}:-r\mathbf{e}\le\mathbf{i}\le(r-1)\mathbf{e}\text{ and }0\le i_1\le r-1\}$;

(iii) *The cardinality of* $P_\varepsilon(r)$ *is* $(2\lfloor\varepsilon r-\frac12\rfloor+1)^{s-1}(\lfloor\varepsilon r-\frac12\rfloor+1)$ *for* $r\ge r_0(\varepsilon)$, *where* $\lfloor a\rfloor$ *denotes the integer part of* a;

(iv) $d^{(\mathbf{i})}:=y_j^{(\mathbf{i})}-x_j$, $1\le j\le N$, *is independent of* j *and*

$$\|d^{(\mathbf{i})}\|\le q\varepsilon,\quad\mathbf{i}\in P_\varepsilon(r),\quad r\ge r_0(\varepsilon).$$

Proof. Assertions (i)–(iii) are straightforward to verify. To prove (iv), note that from Lemma 4.4,

$$d^{(\mathbf{i})}=x_j+\frac{qs^{-1/2}}{r}\left(\mathbf{i}+\frac12\mathbf{e}\right)-x_j=\frac{qs^{-1/2}}{r}\left(\mathbf{i}+\frac12\mathbf{e}\right),$$

which is clearly independent of j. Moreover, as $\mathbf{i}\in P_\varepsilon(r)$, this also implies that

$$\|d^{(\mathbf{i})}\|=\frac{qs^{-1/2}}{r}\|\mathbf{i}+\frac12\mathbf{e}\|\le\frac{qs^{-1/2}}{r}\left(\|\mathbf{i}\|_\infty+\frac12\|\mathbf{e}\|_\infty\right)s^{1/2}\le q\varepsilon.\qquad\square$$

Our primary result can now be stated.

THEOREM 4.6. *Let F belong to RP_0^∞ or RN_1^∞ and let $d\mu$ be its representing measure. Assume that Ω is an admissible domain with respect to a set $X = \{x_1, \dots, x_N\}$ of centers, and that $2q = \min_{j \neq \ell} \|x_j - x_\ell\| > 0$. Suppose that $I'(\mu)$ and $\tilde{J}(\mu, d)$ are given by Theorem* 3.4 *and Remark* 3.5(i), *respectively. If $P_\varepsilon(r)$ and $d^{(\mathbf{i})}$ are as in Lemma* 4.5, *then the following hold:*

(i) *There exists $0 < \varepsilon \le 1$, $\varepsilon = \varepsilon(s, q, \mu, N)$, and $r_0(\varepsilon) \in \mathbb{N}$ such that $P_\varepsilon(r) \neq \emptyset$ for all $r \ge r_0(\varepsilon)$ and*

$$N\tilde{J}(\mu, d^{(\mathbf{i})}) < \tfrac{1}{2} I'(\mu)$$

for $\mathbf{i} \in P_\varepsilon(r)$.

(ii) *With ε as in* (i) *and $\alpha \in \mathbb{R}^N$, $\|\alpha\| = 1$, we have*

$$\langle A_X \alpha, \alpha \rangle \ge 2^{s-2} s^{-s/2} q^s I'(\mu)^2 \varepsilon^s.$$

Proof. (i) It is clear from its definition that $\tilde{J}(\mu, d) \to 0$ with d. So there exists a $\delta = \delta(s, q, \mu, N)$ such that

$$N\tilde{J}(\mu, d) < \tfrac{1}{2} I'(\mu) \quad \text{for } \|d\| \le \delta. \tag{4.4}$$

Choose $\varepsilon = \min(\delta/q, 1)$ and $r_0(\varepsilon)$ as in Lemma 4.5 (i). By assertion (iv) of the same lemma,

$$\|d^{(\mathbf{i})}\| \le q\varepsilon \le \delta, \quad \mathbf{i} \in P_\varepsilon(r), \quad r \ge r_0(\varepsilon).$$

So by (4.4),

$$N\tilde{J}(\mu, d^{(\mathbf{i})}) < \tfrac{1}{2} I'(\mu) \quad \text{for } \mathbf{i} \in P_\varepsilon(r).$$

(ii) By Lemma 4.4,

$$\langle A_X \alpha, \alpha \rangle \ge \underline{\lim}_{r \to \infty} \frac{2q^s s^{-s/2}}{r^s} \sum_{\substack{-r\mathbf{e} \le \mathbf{i} \le (r-1)\mathbf{e} \\ 0 \le i_1 \le r-1}} \|A_1^{(\mathbf{i})}(F)\alpha\|^2.$$

Let ε be chosen as in part (i) and $r \ge r_0(\varepsilon)$. Using the results of part (i) of the present theorem, along with Theorem 3.6 and Lemma 4.5 (ii) and (iii), we obtain

$$\begin{aligned}
\langle A_X \alpha, \alpha \rangle &\ge \underline{\lim}_{r \to \infty} \frac{2q^s s^{-s/2}}{r^s} \sum_{\mathbf{i} \in P_\varepsilon(r)} \|A_1^{(\mathbf{i})}(F)\alpha\|^2 \\
&\ge \underline{\lim}_{r \to \infty} \frac{2q^s s^{-s/2}}{r^s} \left(2 \left\lfloor \varepsilon r - \frac{1}{2} \right\rfloor + 1 \right)^{s-1} \left(\left\lfloor \varepsilon r - \frac{1}{2} \right\rfloor + 1 \right) \frac{I'(\mu)^2}{4} \\
&\ge \underline{\lim}_{r \to \infty} 2^{s-2} s^{-s/2} q^s I'(\mu)^2 \frac{(\lfloor \varepsilon r - \frac{1}{2} \rfloor)^s}{r^s} \\
&= 2^{s-2} s^{-s/2} q^s I'(\mu)^2 \varepsilon^s. \qquad \square
\end{aligned}$$

COROLLARY 4.7. *Suppose that the assumptions of Theorem* 4.6 *hold, and let ε be given by part* (i) *of that theorem. Then the least squares matrix A_X is invertible and*

$$\|A_X^{-1}\| \le 2^{-s+2} s^{s/2} q^{-s} I'(\mu)^{-2} \varepsilon^{-s}.$$

Proof. If $\alpha \in \mathbb{R}^N$ with $\|\alpha\| = 1$, then the Cauchy–Schwarz inequality and Theorem 4.6 imply that $\|A_X\alpha\| \geq 2^{s-2}s^{-s/2}q^s I'(\mu)^2\varepsilon^s$. The required result follows from this because

$$\|A_X^{-1}\| = \left\{\frac{1}{\inf \|A_X\alpha\|} \colon \alpha \in \mathbb{R}^N, \|\alpha\| = 1\right\}. \qquad \square$$

Patently, the quantity ε, in terms of which the estimates in Theorem 4.6 and Corollary 4.7 are given, is of little practical use if it cannot be quantified. Now it is clear from the proof of Theorem 4.6 that the task of quantifying ε is indeed that of describing precisely the relationship between $\tilde{J}(\mu, d)$ and $\|d\|$. The next result serves as an illustration in this regard. Even though it is confined to the three functions F_i, $1 \leq i \leq 3$, the principle of the argument may be extended readily to other functions of one's choosing.

PROPOSITION 4.8. *Suppose that F is one of the three functions F_i, $1 \leq i \leq 3$. Let I_i', $1 \leq i \leq 3$, and J' be given by Propositions* 3.7 *and* 3.8, *respectively. Then the corresponding least squares matrix satisfies the following estimate*:

$$\|A_X^{-1}\| \leq \begin{cases} \dfrac{4s^{s/2}(J')^s N^s}{(I_i')^{s+2}} & \text{if } \dfrac{I_i'}{2J'Nq} < 1; \\[2ex] \dfrac{2^{2-s}s^{s/2}}{q^s(I_i')^2} & \text{if } \dfrac{I_i'}{2J'Nq} \geq 1. \end{cases}$$

Proof. Note that for $F = F_i$, $1 \leq i \leq 3$, we have from Proposition 3.8, $\tilde{J}(\mu, d) \leq J'\|d\|$ (recall here that J' depends only on s). Now let $\delta = I_i'/(2J'N)$ in (4.4), so that ε in Part (i) of Theorem 4.6 may be chosen to be $\min\{I_i'/(2J'Nq), 1\}$. The stated result now follows from Corollary 4.7. $\square$

Remark 4.9. (i) It is important to note that in each of the three estimates (corresponding to F_i, $1 \leq i \leq 3$) given above, the influence of the number of centers N, if present, manifests itself only by way of the factor N^s. However, the influence of the minimal separation distance of the centers varies with the function (as with interpolation). For instance, in the case of F_1, this is reflected in terms of the factor q^{-s-2}, whereas for F_2, the corresponding term is $q^{(3s-s^2)/2}e^{+\text{const}/q}$.

(ii) It is clear from (i) that a small separation parameter q can wreak havoc on the estimates involving the multiquadric. To counter this effect, it is recommended that one use the function $F_2^c(x) := \sqrt{c^2 + \|x\|^2} = c\sqrt{1 + \|x/c\|^2}$, $0 < c < 1$. Note that under this change of scale, the set of centers X with minimal separation distance $2q$ is transformed to a new set Y of centers with minimal separation $2q/c$. Our analysis can now be carried out with respect to the set Y; of course, the new separation parameter is $q' := q/c > q$. In other words, given a set of centers $\{x_i\}$ with (perhaps damagingly) small separation distance, we forgo the basis $F_2(\cdot - x_i)$ in favour of the better conditioned basis $F_2^c(\cdot - x_i)$ for an appropriate choice of c. Clearly, our methods of analysis of F_2^c do not differ significantly from those of F_2 (the transition essentially involves a mere change in scale). Yet, this choice of a new basis will improve considerably the overall stability of the least squares process.

To explain further this improvement in stability, let us draw attention to the fact that the two limiting cases of F_2^c, viz., $c = 0$ and $c = 1$, correspond to the functions F_1 and F_2, respectively. Recalling the roles of the separation parameter q in the estimates involving these limiting cases, we see that the exponential influence of q, felt in the case of $c = 1$, gets mitigated for smaller values of c as the respective bounds gradually begin to mirror the estimate corresponding to $c = 0$.

(iii) It is possible to obtain an estimate for the least squares matrix associated with the function $F(x) = \|x\|^{2\beta}$, $x \in \mathbb{R}^s$, $0 < \beta < 1$, using Remark 3.10 (i) and Corollary 4.7.

(iv) The estimate for F_2 derived in Proposition 4.8 can be improved using Remark 3.10 (ii); this is because $\tilde{J}(\mu, d)$ can actually be estimated in terms of $\|d\|^2$. Precisely, the following upper bound may be obtained:

$$\|A_X^{-1}\| \leq \begin{cases} \dfrac{2^{-s/2+2}s^{s/2}(J'')^{s/2}N^{s/2}}{(I_2')^{s/2+2}} & \text{if } \left(\dfrac{I_2'}{2J''Nq^2}\right)^{1/2} < 1; \\ \dfrac{2^{2-s}s^{s/2}}{q^s(I_2')^2} & \text{if } \left(\dfrac{I_2'}{2J''Nq^2}\right)^{1/2} \geq 1. \end{cases}$$

It is evident that the estimates in Theorem 4.6 (and hence those of Corollary 4.7 and Proposition 4.8) all involve the number of centers N. It is no less evident that the entry of this factor was occasioned solely by the restriction $N\tilde{J}(\mu, d^{(\mathbf{i})}) < I'(\mu)/2$ - an unavoidable consequence of the interpolation estimates of Theorem 3.6. However, as the interpolation estimate for the Gaussian given in Proposition 3.11 avoids the quantity N, it is not surprising that the following least squares estimate for F_4 also shares the same desirable feature.

PROPOSITION 4.10. *Let I_4' and $D_s(\rho, q)$ be as in Propositions 3.7 and 3.11, respectively. Suppose that A_X is the least squares matrix associated with F_4. Then*

$$\|A_X^{-1}\| \leq \begin{cases} \dfrac{4s^{s/2}[D_s(\rho,q)]^s}{(I_4')^{s+2}} & \textit{if } \dfrac{I_4'}{2D_s(\rho,q)q} < 1; \\ \dfrac{2^{2-s}s^{s/2}}{q^s(I_4')^2} & \textit{if } \dfrac{I_4'}{2D_s(\rho,q)q} \geq 1. \end{cases}$$

Proof. We proceed as in Theorem 4.6, also adhering to the notation therein. Let $\alpha \in \mathbb{R}^N$, $\|\alpha\| = 1$. As before, we have

$$\langle A_X\alpha, \alpha\rangle \geq \underline{\lim}_{r\to\infty} \frac{2q^s s^{-s/2}}{r^s} \sum_{\substack{-r\mathbf{e} \leq \mathbf{i} \leq (r-1)\mathbf{e} \\ 0 \leq i_1 \leq r-1}} \|A_1^{(\mathbf{i})}(F_4)\alpha\|^2.$$

We wish to bound $\|A_1^{(\mathbf{i})}(F_4)\alpha\|^2$ from below without involving N. To this end, we invoke Proposition 3.11: If $\varepsilon = \min\{I_4'/(2D_s(\rho,q)q), 1\}$ and $\mathbf{i} \in P_\varepsilon(r)$, then $\|d^{(\mathbf{i})}\| \leq q\varepsilon \leq q$ and $D_s(\rho,q)\|d^{(\mathbf{i})}\| < I_4'/2$. Therefore, as in the proof of part (ii) of Theorem 4.6, we have

$$\langle A_X\alpha, \alpha\rangle \geq 2^{s-2}s^{-s/2}q^s(I_4')^2\varepsilon^s.$$

This implies

$$\langle A_X\alpha, \alpha\rangle \geq \begin{cases} \dfrac{(I_4')^{s+2}}{4s^{s/2}[D_s(\rho,q)]^s} & \text{if } \dfrac{I_4'}{2D_s(\rho,q)q} < 1; \\ \dfrac{2^{s-2}(I_4')^2q^s}{s^{s/2}} & \text{if } \dfrac{I_4'}{2D_s(\rho,q)q} \geq 1, \end{cases}$$

whence the required result follows as in Corollary 4.7. □

Remark 4.11. Recall that $D_s(\rho,q) = \pi^{-s}\rho q\left[1 + 6s\sum_{n=0}^{\infty}(n+3)^s e^{-\rho n^2 q^2}\right]$ and $I_4' = C_s e^{-\delta^2/(q^2\rho)}/(q^s\rho^{s/2})$. So it is obvious from the preceding proposition that the number N plays no role whatsoever in determining the stability of A_X^{-1}. On the other hand, as in the case of the multiquadric, a small separation parameter q can have a severe impact on the estimate. The scaling parameter ρ can now be chosen judiciously to alleviate this negative effect. For instance, if ρ is chosen to be of the order q^{-2}, then both I_4' and the term $e^{-\rho n^2 q^2}$ are rendered independent of q, and $D_s(\rho, q)$ is of the order q^{-1}. Thus, the impact (on the least squares estimate) of the exponential term involving q can be lessened significantly to that of q^{-s}.

5. Some sharper estimates. If $s = 1$, $F = F_1$, and the centers are equally spaced (i.e., $q = \text{const}/N$) in some admissible domain, then Proposition 4.8 and Remark 4.9 (i) indicate that the norm of the inverse of the associated least squares matrix is $O(N^4)$. We shall demonstrate that it is actually $O(N^3)$. To this end, a careful scrutiny of the arguments leading up to Theorem 4.6, and consequently Proposition 4.8, reveals that the quantity N^3 in the estimate for $\langle A_X\alpha, \alpha\rangle$—and hence for $\|A_X^{-1}\|$—comes from $(I_1')^{s+2}$, whereas the extra factor N is a result of the restriction $N\tilde{J}_1 < I_1'/2$ in Theorem 4.6. If, however, we can somehow bypass $\tilde{J}_1$, and, as with the Gaussian, estimate $J_1/\|\alpha\|^2$ (where $J_1 := J(\mu_1, d)$) directly by a quantity that is independent of N (but approaching zero with d, of course), then the desired sharper estimate will obtain. So let us consider the expression J_1 for $s = 1$. Recalling that $F_1 \in RN_1^\infty$, we find from Theorem 3.2 that for $s = 1$,

$$J_1 = \frac{1}{\pi}\int_0^\infty \frac{1}{t^{3/2}}\int_{\mathbb{R}} e^{-u^2/4t}\sin^2\left(\frac{ud}{2}\right)\left|\sum_{j=1}^N \alpha_j e^{i\cdot ux_j}\right|^2 du\,d\mu_1(t)$$
$$= \frac{1}{2\pi^{3/2}}\int_{\mathbb{R}} \sin^2\left(\frac{ud}{2}\right)\left|\sum_{j=1}^N \alpha_j e^{iux_j}\right|^2\left[\int_0^\infty e^{-u^2/4t}\frac{dt}{t^2}\right]du,$$

by the definition of $d\mu_1(t)$ and Fubini's theorem. Evaluating the inner integral using the substitution $w = 1/4t$, we see that

$$J_1 = \frac{2}{\pi^{3/2}}\int_{\mathbb{R}} \frac{\sin^2(ud/2)}{u^2}\left|\sum_{j=1}^N \alpha_j e^{iux_j}\right|^2 du$$
$$= \frac{d^2}{2\pi^{3/2}}\int_{\mathbb{R}} \left(\frac{\sin(ud/2)}{ud/2}\right)^2\left|\sum_{j=1}^N \alpha_j e^{iux_j}\right|^2 du$$
$$= \frac{d^2}{2\pi^{3/2}}\sum_{j,\ell=1}^N \alpha_j\alpha_\ell\int_{\mathbb{R}} \left(\frac{\sin(ud/2)}{ud/2}\right)^2 e^{iu(x_j - x_\ell)}du.$$

Let $M(x) := (1-|x|)\chi_{[-1,1]}(x)$ denote the centered second-order cardinal B-spline. Recall that $\widehat{M}(\xi) = (\sin(\xi/2)/(\xi/2))^2, \xi \in \mathbb{R}$, so

$$J_1 = \frac{d^2}{2\pi^{3/2}}\sum_{j,\ell=1}^N \alpha_j\alpha_\ell\int_{\mathbb{R}} \widehat{M}(ud)e^{iu(x_j - x_\ell)}du$$
$$= \frac{d}{2\pi^{3/2}}\sum_{j,\ell=1}^N \alpha_j\alpha_\ell\int_{\mathbb{R}} [M(\cdot/d)]^\wedge(u)e^{iu(x_j - x_\ell)}du$$
$$= \frac{d}{\sqrt{\pi}}\sum_{j,\ell=1}^N \alpha_j\alpha_\ell M\left(\frac{x_j - x_\ell}{d}\right).$$

Now, if $d < 2q$ (remember that we are interested only in small d), then $|x_j - x_\ell| > d$ for $j \neq \ell$, so $M\left(x_j - x_\ell/d\right) = 0$, $j \neq \ell$. Consequently,

$$J_1 = \frac{d}{\sqrt{\pi}}\|\alpha\|^2, \qquad d < 2q,$$

which is an expression clearly free of N. (We also note in passing that since $d\mu_2(t) \leq d\mu_1(t)$,

$$J_2 := J(\mu_2, d) \leq J_1 = \frac{d}{\sqrt{\pi}}\|\alpha\|^2 \text{ provided } s = 1 \quad \text{and} \quad d < 2q.$$

A similar statement also holds for $J(\mu_3, d)$. Consequently, the estimate for $\|A_X^{-1}\|$, corresponding to $F = F_2$ and F_3 for equally spaced centers, can be strengthened in the univariate case.)

Having derived the estimate $\|A_X^{-1}\| = O(N^3)$ for $F = F_1$, $s = 1$, and equally spaced centers, we now wish to point out that it is optimal at least when the number of centers is odd. Let the centers x_i be given by $x_i := i/(k+1) \in \Omega = [0,1]$, $i = 1, \ldots, k = 2m+1$. We shall show that there exists a constant D such that $\|A_X^{-1}\| \geq Dk^3$ for large k.

Firstly, a direct computation yields an exact expression for the entries of A_X:

$$\begin{aligned} A_X(i,j) = \langle |\cdot - x_i|, |\cdot - x_j|\rangle &= \frac{1}{3} - \frac{(i+j)}{2(2m+2)} + \frac{ij}{(2m+2)^2} + \frac{|i-j|^3}{3(2m+2)^3} \\ &=: P_{ij} - \frac{1}{3(2m+2)^3} G(|i-j|), \end{aligned}$$

where $G(x) = -|x|^3$. Now we select suitable vectors λ which "annihilate" linear polynomials, and use them to estimate $\|A_X\lambda\|/\|\lambda\|$ from above. Indeed, let $\lambda \in \mathbb{R}^{2m+1}$ be given by

$$\lambda_j := (-1)^j \binom{2m}{j}, \qquad 0 \leq j \leq 2m.$$

If P denotes the matrix whose entries are P_{ij} (as defined above), then $P\lambda = 0$. Thus,

$$A_X\lambda = \frac{-1}{3(2m+2)^3} B\lambda, \quad \text{where } B_{ij} = G(|i-j|).$$

Clearly, B corresponds to the radial interpolation matrix associated with G. An appeal to Theorems 3.4 and 4.1 of [BSW] (taking q there to be 1) reveals that

$$\frac{\|B\lambda\|^2}{\|\lambda\|^2} = O(1), \qquad m \to \infty.$$

Thus,

$$\|A_X^{-1}\| \geq D(2m+2)^3 \text{ for large } m,$$

and the proof is complete.

(We do note that in [BSW], the aforesaid assertion was actually proved for the function $x \mapsto |x|^{2\beta}$, $0 < \beta < 1$, which belongs to RN_1^∞. The function $G(x) = -|x|^3$, however, belongs to RN_2^∞, but the proof in [BSW] does extend to the case $\beta = \frac{3}{2}$ and $m = 2$. The representing measure for G is given by $\left(\frac{3}{4}\right)(\pi t)^{-1/2}\,dt$.)

Acknowledgment. We thank Phil Smith for a stimulating conversation concerning least squares approximation.

REFERENCES

[B] K. BALL, *Eigenvalues of Euclidean distance matrices*, J. Approx. Theory, 68 (1992), pp. 74–82.

[de B] C. DE BOOR, *A practical guide to splines*, Springer-Verlag, Berlin, New York, 1978.

[Ba] B. J. C. BAXTER, *Norm estimates for inverses of Toeplitz distance matrices*, DAMTP 1991/NA16, J. Approx. Theory, to appear.

[BDR] C. DE BOOR, R. DEVORE, AND A. RON, *Approximation from shift-invariant subspaces of* $L_2(\mathbb{R}^d)$, TSR#92-02, University of Wisconsin, Madison, WI, 1991, Trans. Amer Math. Soc., to appear.

[BSW] K. BALL, N. SIVAKUMAR, AND J. D. WARD, *On the sensitivity of radial basis interpolation to minimal data separation distance*, Constr. Approx., 8 (1992), pp. 401–426.

[C] E. W. CHENEY, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.

[D] N. DYN, *Interpolation and approximation by radial and related functions*, Approximation Theory VI, Vol. 1, C. K. Chui, L. L. Schumaker, and J. D. Ward, eds., Academic Press, New York, 1989, pp. 211–234.

[GV] I. M. GELFAND AND N. YA. VILENKIN, *Generalized Functions*, Vol. 4, Academic Press, New York, 1965.

[M] C. A. MICCHELLI, *Interpolation of scattered data: distances, matrices, and conditionally positive definite functions*, Constr. Approx., 2 (1986), pp. 11–22.

[MN] W. R. MADYCH AND S. A. NELSON, *Multivariate interpolation: a variational theory*, manuscript, 1983.

[NW1] F. J. NARCOWICH AND J. D. WARD, *Norms of inverses and condition numbers for matrices associated with scattered data*, J. Approx. Theory, 64 (1991), pp. 69–94.

[NW2] ———, *Norm estimates for the inverses of a general class of scattered-data radial-basis interpolation matrices*, J. Approx. Theory, 69 (1992), pp. 84–109.

[P] M. J. D. POWELL, *The theory of radial basis function approximation*, in Wavelets, Subdivision, and Radial functions, W. Light, ed., Oxford University Press, London, 1990.

[S1] I. J. SCHOENBERG, *On certain metric spaces arising from Euclidean space by a change of metric and their imbedding in Hilbert space*, Ann. Math., 38 (1937), pp. 787–793.

[S2] ———, *Metric spaces and completely monotone functions*, Ann. Math., 39 (1938), pp. 811–841.

[Su1] X. SUN, *On the solvability of radial function interpolation*, Approximation Theory VI, Vol. 2, C.K. Chui, L.L. Schumaker, and J.D. Ward, eds., Academic Press, New York, 1989, pp. 643–646.

[Su2] ———, *Norm estimates for inverses of Euclidean distance matrices*, J. Approx. Theory, 70 (1992), pp. 339–347.

SIAM J. MATH. ANAL.
Vol. 24, No. 4, pp. 1067–1085, July 1993

ESTIMATES FOR LARGE DEVIATIONS IN RANDOM TRIGONOMETRIC POLYNOMIALS*

GEORGE BENKE[†] [‡] AND W. J. HENDRICKS[†]

Abstract. Let $F(t) = \sum_{n=1}^{N} a_n \exp(iX_n t)$, where $X_1, X_2, \ldots, X_N$ are independent random variables and the coefficients a_n are real or complex constants. Probabilistic estimates of the form

$$P\left[\sup_{t\in K} |F(t) - E[F(t)]| \geq C\sqrt{N\log N}\right] \leq \epsilon$$

are obtained where K is an interval on the real line, C may be chosen more or less arbitrarily, and ϵ is an explicit function of C, K, N, and the random variables. This method includes trigonmetric interpolation and straightforward probabilistic techniques to obtain explicit numerical bounds that are applicable in a variety of engineering applications, particularly in the study of maximal sidelobe level for random arrays. Specific numerical examples are computed, and references to both the engineering and mathematical literature are provided.

Key words. large deviations, random trigonometric polynomial, random array, maximum sidelobe level

AMS subject classifications. primary 60F10, 60G17, 60G35; secondary 42A05

1. Introduction.

In this paper we obtain probabilistic bounds for large deviations for a particular class of random trigonometric polynomials and show that these estimates are associated in a natural way with what is known in the engineering literature as the maximal sidelobe level problem for random phased arrays. These arrays have been studied extensively in the engineering community through simulations and various approximations, but a mathematically rigorous presentation of their probabilistic properties has not been given in either the mathematical or the engineering literature. Since the problem embraces both the engineering and mathematical disciplines, the subsequent paragraphs of this introductory section provide an explanation of the relevant terminology and of the mathematical and physical setting for the problem. In §2 we develop the means for deriving the probabilistic estimates through a sequence of lemmas, after which we are able to state our main theorem and ensuing corollaries. Section 3 gives specific examples of these results.

A phased array consists of a finite collection of transmitting or receiving elements distributed in space. These elements transmit or receive energy from some field such as an acoustic, electromagnetic, or seismic field. By adjusting the phases of the signals at the individual sensing elements of the array the contribution from each element can be added coherently with those of the other array elements, resulting in a more sensitive array response to an incoming signal. The array elements, along with the inter-element connections required to coherently add the elemental responses, constitute a phased array antenna. Antennas of this type have been used for many years in areas such as radar, sonar, seismology, and radio astronomy. The books by Steinberg [15] and Haykin [5] give a more detailed overview of these application areas. One advantage of

*Received by the editors June 9, 1992; accepted for publication (in revised form) August 31, 1992.

[†] The MITRE Corporation, 7525 Colshire Drive, McLean, Virginia 22102.

[‡] Mathematics Department, Georgetown University, Washington, D.C. 20057.

phased array antennas derives from the fact that array elements can be located over a large region in space. This gives the antenna a large effective aperture, and therefore the ability to form very narrowly focused beams. By adjusting the phases of the signals at the individual elements, the beam of the antenna can be steered quickly in different directions without the motion of any physical components. Another advantage results from the fact that the signals from each sensor are accessible, thereby allowing for sophisticated processing algorithms that can mitigate the effects of correlated noise, multipath, and other undesirable phenomena.

In order to describe more fully the notion of a phased array, consider an array of omnidirectional receiving elements located at positions $\mathbf{x}_1, \dots, \mathbf{x}_N$ in Euclidean space $\mathbf{R}^3$, and let S be a propagating scalar plane wave with waveform s. That is, let s be a complex-valued function defined on the real line $\mathbf{R}$ and define S at position $\mathbf{x} \in \mathbf{R}^3$ and time $\tau \in \mathbf{R}$ by

$$S(\mathbf{x}, \tau) = s\left(\tau - \frac{\mathbf{n} \cdot \mathbf{x}}{c}\right),$$

where $\mathbf{n}$ is a unit vector indicating the direction of propagation and c is the propagation velocity of the wave. The element at position $\mathbf{x}_n$ receives the signal

$$g_n(\tau) \equiv S(\mathbf{x}_n, \tau) + \eta_n(\tau),$$

where η_n is a sample function of a stochastic process that models the noise at time τ and position $\mathbf{x}_n$. Since the array elements are located at different positions, radiation from a given direction will typically arrive at different sensors at different times. As a result, a delay must be applied at each sensor if the elemental signal responses are to add coherently. In "delay and sum" processing, the functions g_n are combined in this way by forming

$$D(\tau) \equiv \sum_{n=1}^{N} g_n(\tau - \tau_n) = \sum_{n=1}^{N} s\left(\tau - \tau_n - \frac{\mathbf{n} \cdot \mathbf{x}_n}{c}\right) + \sum_{n=1}^{N} \eta_n(\tau - \tau_n),$$

where $\tau_1, \dots, \tau_N$ represent a set of delays. If we choose $\tau_n = -(\mathbf{n} \cdot \mathbf{x}_n)/c$, then

$$D(\tau) = Ns(\tau) + H(\tau),$$

where H gives the combined noise output of the array, and we say that we have steered the beam of the array in the direction $\mathbf{n}$. Note that this choice of the τ_n gives an array signal response of N times that of each of the elemental signal responses. By adjusting the delays we thus observe that the maximal array response can be made to point in any prescribed direction—hence the notion of steering the beam.

Without being overly concerned about the existence of limits, let us define

$$\langle g, h \rangle = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} g(t)\overline{h(t)}dt \quad \text{and} \quad \|h\| = \langle h, h \rangle^{1/2}.$$

Assume that the noise components and signal are uncorrelated, that is, assume that

$$\langle \eta_k, \eta_j \rangle = 0 \quad \text{for } k \neq j \quad \text{and} \quad \langle \eta_n, s \rangle = 0 \quad \text{for } n = 1, \dots, N.$$

Assume also that $\|\eta_n\|$ is the same for all n, and, therefore, denote $\|\eta_n\|$ by $\|\eta\|$. Then:

$$\left(\frac{\|D\|}{\|H\|}\right)^2 = N\left(\frac{\|s\|}{\|\eta\|}\right)^2 + 1;$$

therefore, the (power) signal to noise ratio for the array is approximately N times the signal to noise ratio at each receiving element. Thus, arrays with arbitrarily high sensitivity can be created by using sufficiently many sensors.

In order to examine the directional sensitivity of an array it is convenient to consider a monochromatic signal of (radian) frequency ω:

$$s(\tau) = \exp(i\,\omega\,\tau),$$

and to neglect noise. Thus if $\mathbf{n}$ gives the direction of a propagating plane wave

$$S(\mathbf{x},\tau) = \exp\left(i\omega\left(\tau - \frac{\mathbf{n}\cdot\mathbf{x}}{c}\right)\right)$$

and delays τ_n are chosen to maximize sensitivity in a direction $\mathbf{n}_0$, the delay and sum array output is

$$\begin{aligned} D(\tau) &= \sum_{n=1}^{N} \exp\left(i\omega\left(\tau + \frac{\mathbf{n}_0\cdot\mathbf{x}_n}{c} - \frac{\mathbf{n}\cdot\mathbf{x}_n}{c}\right)\right) \\ &= e^{i\omega\tau}\sum_{n=1}^{N}\exp\left(i\frac{\omega}{c}(\mathbf{n}_0 - \mathbf{n})\cdot\mathbf{x}_n\right) \\ &= s(\tau)F\left(\frac{\omega}{c}(\mathbf{n}_0 - \mathbf{n})\right), \end{aligned}$$

where F is the trigonometric polynomial on $\mathbf{R}^3$ given by

$$F(\boldsymbol{\gamma}) = \sum_{n=1}^{N}\exp\left(i\boldsymbol{\gamma}\cdot\mathbf{x}_n\right).$$

Note that the array output D can be expressed as the product of a time factor s (the signal) and a space factor F, and that $F(0) = N$. Infinite spatial resolution would be obtained if $F(\boldsymbol{\gamma}) = 0$ for all $\boldsymbol{\gamma} \neq 0$—i.e., if F were a delta function. The directional characteristics of the array are, therefore, embodied in the trigonometric polynomial F. The question then becomes "How closely does F resemble a delta function?" Being a trigonometric polynomial, F is almost periodic, so that, strictly speaking, F is globally never very close to a delta function. However, since $|\mathbf{n} - \mathbf{n}_0| \leq 2$, the array directivity is determined for all directions by $F(\boldsymbol{\gamma})$ for $|\boldsymbol{\gamma}| \leq 2\omega/c$. This ball is called the visible region. Recall that the Fourier transform of the delta function is a constant and the trigonometric polynomial F is (up to normalization) the inverse Fourier transform of the discrete measure μ with unit masses at $\mathbf{x}_1, \ldots, \mathbf{x}_N$. The problem of approximating the delta function by F is, therefore, equivalent to approximating (in some sense) Lebesgue measure by the discrete measure μ. It seems reasonable, therefore, to space the $\mathbf{x}_n$ evenly over some region of space (the aperture). However, with uniform spacing of array elements F becomes periodic with a period that varies linearly with N/L, where L is the size of the aperture. Hence, if N is fixed, for sufficiently large apertures there will exist $\boldsymbol{\gamma}_0 \neq 0$ in the visible region for which $F(\boldsymbol{\gamma}_0) = N$, and the graph of $|F|$ will have a large peak at $\boldsymbol{\gamma}_0$. Such peaks are called grating lobes. These are undesirable since they give the antenna a multiple beam characteristic. On the other hand, large apertures cause $|F(\boldsymbol{\gamma})|$ to fall rapidly from peaks where $F(\boldsymbol{\gamma}) = N$. This is a desirable feature, since it gives the antenna array a narrow main beam, thereby providing good spatial resolution. It is also desirable to keep N, the number of array sensors, small relative to the aperture size L since the cost of an array is largely determined by the number of sensor elements.

In the 1960s Lo, in a series of papers [11]–[13], introduced an approach for overcoming the grating lobe problem while keeping the element density N/L small. Lo's method was to space the array elements randomly, thereby destroying the periodicities that occur when array elements are spaced uniformly. The array is then characterized by the random trigonometric polynomial

$$F(\boldsymbol{\gamma}) = \sum_{n=1}^{N} \exp\left(i\boldsymbol{\gamma}\cdot\mathbf{X}_n\right),$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are independent random variables. The maximal sidelobe level problem in random array theory is to determine the distribution of the random variable

$$\sup_{\boldsymbol{\gamma}\in K} |F(\boldsymbol{\gamma})|$$

for a given compact set K. This is an extremely difficult problem. However, for most practical scenarios the probability density functions of the $\mathbf{X}_n$ are such that the expectation $E[F(\boldsymbol{\gamma})]$ can be explicity computed for each $\boldsymbol{\gamma}$. It is, therefore, of interest to estimate probabilities of the following type:

$$P\left[\sup_{\boldsymbol{\gamma}\in K} |\, F(\boldsymbol{\gamma}) - E[F(\boldsymbol{\gamma})]\,| > \lambda\right]$$

for various (nonnegative) values of λ.

The particular problem considered in this paper is the one-dimensional problem of estimating the supremum of the modulus of expressions of the form

$$G(t) = \sum_{n=1}^{N} a_n \left(e^{iX_n t} - E\left[e^{iX_n t}\right]\right), \tag{1.1}$$

where $X_1, X_2, \ldots, X_N$ are independent real-valued random variables that assume values in the interval $[-L, L]$, $a_1, a_2, \ldots, a_N$ are real or complex constants, and t is a real number. In the context of one-dimensional line arrays, the variable t corresponds to the direction variable $(\omega/c)\cos\theta$ (not time), where θ is the angle between the propagation direction and the line containing the array elements. Consequently, the set K can be regarded as representing an interval of arrival angles, and the problem being considered is that of estimating the peak response to signals emanating from this set of arrival angles. The engineering literature (e.g., Donvito and Kassam [3], Lo [11]–[13], Steinberg [15], [16]) contains several interesting studies of this problem, along with accompanying computer simulations of random sums of the form (1.1). Frequently there are simplifying assumptions such as process stationarity or an omission of a rigorous analysis of the effects of Central Limit Theorem approximations. Such simplifications are understandable, due to the difficulty of obtaining precise probabilistic estimates for these sums.

A different type of random trigonometric polynomial has been the object of mathematical investigation by Kahane [8], who considers the supremum problem for series of the form

$$F(t) = \sum_{n=1}^{N} X_n f_n(t), \tag{1.2}$$

where the f_n are complex trigonometric polynomials and the X_n are various types of random variables. In engineering terminology, Kahane considers signals having random amplitudes and deterministic frequencies, while we consider signals having deterministic amplitudes and random frequencies. Kahane (Chapter 6) traces the study of random trigonometric polynomials of the form (1.2) back to Zygmund [18], who considers convergence properties of such series ($N = \infty$), and to Salem and Zygmund [14]. Observe that in the work of Kahane and his predecessors the frequencies in the trigonometric polynomials are integral, resulting in periodic functions. By contrast, the frequencies X_n that we consider are nonintegral, so that our random functions are not periodic. This creates a significant problem in the subsequent analysis.

For $G(t)$ given by (1.1), we will obtain estimates of the form

$$P\left[\sup_{t\in K} |G(t)| \geq C\sqrt{N\log N}\right] \leq \epsilon,$$

where K is some interval on the real line. The constant C may be chosen more or less arbitrarily, and the ϵ depends on the values of C, K, N, and the properties of the random variables that specify the frequencies. In Kahane's analysis of the bounds for the distribution of the supremum over the set K, any dependency upon the measure of the set K is absorbed in unspecified constants occurring in his estimates. Our formulation, which is oriented toward explicit estimates that might occur in an application, exhibits the specific role that the measure of K and the other parameters play in the estimates.

In the analysis of the probability of large deviations of $|G(t)|$ it is tempting to use Central Limit Theorem approximations because the resulting quantities lend themselves to relatively straightforward analysis. It should be observed, however, that unless there is a careful error analysis when Central Limit Theorem approximations are invoked there may be probabilistic terms on the order of $1/\sqrt{N}$ that must be included in whatever estimates are proposed. For example, see the discussion on the Berry-Esseen error estimate that is presented in detail in Chung [2, §7.1], Breiman [1, p. 184], or Feller [4, p. 515]. In particular, it should be noted that Chung [2, exercise 3, p. 231] gives an example of a $1/\sqrt{N}$ *lower* bound for the Central Limit Theorem error approximation. As a consequence, when the probabilities we seek to estimate are less than $1/\sqrt{N}$, a more sensitive analysis is required.

It should be emphasized at the outset that we are especially interested in obtaining probabilistic estimates for large deviations of $|G(t)|$ when N, the number of summands, is large, perhaps on the order of 10,000. Hence, the probabilities being estimated are quite small. We make no assumptions about stationarity of the process $G(t)$, and do not use any Central Limit Theorem approximations to estimate the probabilities of interest. Any approximations made during the course of our analysis are incorporated explicitly in the estimates that we provide. Despite the mathematical rigors that this approach imposes, we are able to provide reasonable bounds that are useful in applications when the number of summands is large, along with the assurance that there are no terms unaccounted for that might dominate the probabilities being estimated.

2. Main results. In this section we present our main theorem and several simple corollaries. While the theorem is fairly general in its statement, the corollaries are sufficiently specific to lead directly to applications. We illustrate the use of these corollaries and the main theorem in the following section. The proof of the main theorem is broken down into a sequence of lemmas, the first of which is of some interest in its own right. It is a Shannon–Whittaker type of sampling theorem where the interpolation formula involves an absolutely convergent series whose convergence rate is governed by a parameter R. This same parameter also governs the degree of oversampling required as compared to the usual Nyquist rate.

We begin with some notation and definitions (see [9]). Let K denote the function

$$K(t) = \frac{1}{2\pi}\left(\frac{\sin t/2}{t/2}\right)^2 .$$

Then the Fourier transform $\widehat{K}$ is given by

$$\widehat{K}(\gamma) = \int_{-\infty}^{\infty} K(t)\, e^{-i\gamma t}\, dt = \max\{1 - |\gamma|, 0\}.$$

The Fourier inversion theorem also applies, giving

$$K(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{K}(\gamma)\, e^{i\gamma t}\, d\gamma.$$

For $R > 1$ denote by V_R the function

$$V_R(t) = \frac{1}{R-1}\left(R^2K(Rt) - K(t)\right).$$

Then

$$\widehat{V}_R(\gamma) = \frac{R}{R-1}\widehat{K}\left(\frac{\gamma}{R}\right) - \frac{1}{R-1}\widehat{K}(\gamma).$$

Thus, $\widehat{V}_R$ is real valued and has a graph that is a trapezoid extending from $-R$ to R with a flat section of height 1 extending from -1 to $+1$.

LEMMA 1. *Let ν be a bounded measure supported in the interval $[-L, L]$, and let*

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{it\gamma}\, d\nu(\gamma)$$

denote the inverse Fourier transform of ν. Then for any $R > 1$ and all real t

$$f(t) = \frac{\pi}{R}\sum_{n=-\infty}^{\infty} V_R\left(Lt - \frac{n\pi}{R}\right) f\left(\frac{n\pi}{RL}\right),$$

where the series converges absolutely for all t and

$$V_R(t) = \frac{1}{2\pi(R-1)}\left(R^2\left(\frac{\sin Rt/2}{Rt/2}\right)^2 - \left(\frac{\sin t/2}{t/2}\right)^2\right).$$

Proof. Let $V_{R,L}(t) = L\,V_R(Lt)$; then

$$\widehat{V}_{R,L}(\gamma) = \begin{cases} 1 & \text{for } |\gamma| \le L, \\ 0 & \text{for } |\gamma| \ge RL, \end{cases}$$

and is continuous and piecewise linear. For fixed $t \in \mathbf{R}$, let $\phi_t(\gamma) = \widehat{V}_{R,L}(\gamma)\,e^{i\gamma t}$. Expressing ϕ_t in a Fourier series relative to the interval $[-RL, RL]$ gives

$$\phi_t(\gamma) = \sum_{n=-\infty}^{\infty} c_n(t)\exp\left(i\frac{\pi n\gamma}{RL}\right), \quad \text{where} \tag{2.1}$$

$$\begin{aligned} c_n(t) &= \frac{1}{2RL}\int_{-RL}^{RL}\phi_t(\xi)\exp\left(-\frac{i\pi n\xi}{RL}\right)d\xi \\ &= \frac{\pi}{RL}\frac{1}{2\pi}\int_{-\infty}^{\infty}\widehat{V}_{R,L}(\xi)\exp\left(i\left(t-\frac{n\pi}{RL}\right)\xi\right)d\xi \\ &= \frac{\pi}{RL}V_{R,L}\left(t-\frac{n\pi}{RL}\right) \\ &= \frac{\pi}{R}V_R\left(Lt-\frac{n\pi}{R}\right). \end{aligned}$$

When this expression is substituted into (2.1) we obtain

$$\phi_t(\gamma) = \frac{\pi}{R}\sum_{n=-\infty}^{\infty} V_R\left(Lt-\frac{n\pi}{R}\right)\exp\left(i\frac{\pi n\gamma}{RL}\right), \tag{2.2}$$

where the series converges absolutely.

Now consider $f(t) = (1/2\pi)\int_{-L}^{L} e^{i\gamma t}\,d\nu(\gamma)$. Since ν is supported in $[-L, L]$, we may write

$$f(t) = \frac{1}{2\pi}\int_{-RL}^{RL}\widehat{V}_{R,L}(\gamma)\,e^{i\gamma t}\,d\nu(\gamma) = \frac{1}{2\pi}\int_{-RL}^{RL}\phi_t(\gamma)\,d\nu(\gamma).$$

Using the absolute convergence of the series (2.1) gives

$$\begin{aligned} f(t) &= \frac{1}{2\pi}\int_{-RL}^{RL}\left(\frac{\pi}{R}\sum_{n=-\infty}^{\infty}V_R\left(Lt-\frac{n\pi}{R}\right)\exp\left(i\frac{\pi n\gamma}{RL}\right)\right)d\nu(\gamma) \\ &= \frac{\pi}{R}\sum_{n=-\infty}^{\infty}V_R\left(Lt-\frac{n\pi}{R}\right)\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp\left(i\frac{\pi n\gamma}{RL}\right)d\nu(\gamma) \\ &= \frac{\pi}{R}\sum_{n=-\infty}^{\infty}V_R\left(Lt-\frac{n\pi}{R}\right)f\left(\frac{n\pi}{RL}\right). \qquad \square \end{aligned}$$

LEMMA 2. *For all $y \in \mathbf{R}$,*

$$\frac{\pi}{R}\sum_{n=-\infty}^{\infty}\left|V_R\left(y-\frac{n\pi}{R}\right)\right| \le \frac{R+1}{R-1}, \quad \textit{where}$$

$$V_R(t)=\frac{1}{2\pi(R-1)}\left(R^2\left(\frac{\sin Rt/2}{Rt/2}\right)^2-\left(\frac{\sin t/2}{t/2}\right)^2\right).$$

Proof. We use the Poisson summation formula, which is

$$2\pi\lambda\sum_{n=-\infty}^{\infty} f(2\pi\lambda n)=\sum_{n=-\infty}^{\infty}\widehat{f}\left(\frac{n}{\lambda}\right), \quad \text{where}$$

$\widehat{f}(\gamma)=\int_{-\infty}^{\infty} f(t)\,e^{-i\gamma t}\,dt$ and λ is any positive constant. This formula is valid for $f\in L^1(\mathbf{R})$, if f is also of bounded variation (Zygmund [18, p. 68]).

Let $f(x)=K(x-y)$, so that $\widehat{f}(\gamma)=e^{-i\gamma y}\,\widehat{K}(\gamma)$. Since K is in $L^1(\mathbf{R})$ and is also of bounded variation, so is f. Using the Poisson summation formula with $\lambda=\frac{1}{2}R$ gives

$$\begin{aligned}\frac{\pi}{R}\sum_{n=-\infty}^{\infty}K\left(y-\frac{n\pi}{R}\right)&=2\pi\lambda\sum_{n=-\infty}^{\infty}f(2\pi\lambda n)\\&=\sum_{n=-\infty}^{\infty}\widehat{f}(2Rn)=\sum_{n=-\infty}^{\infty}e^{-i2Rny}\,\widehat{K}(2Rn).\end{aligned}$$

Since $\widehat{K}(\gamma)=0$ for $|\gamma|\ge 1$, the right-hand sum has only one nonzero term, corresponding to $n=0$. Inasmuch as $\widehat{K}(0)=1$, this term evaluates to 1. The same argument applied to $R\,K\,(Rx)$ instead of $K(x)$ gives

$$\frac{\pi}{R}\sum_{n=-\infty}^{\infty}RK\left(R\left(y-\frac{n\pi}{R}\right)\right)=1 \quad \text{for all } y,$$

whence

$$\begin{aligned}&\frac{\pi}{R}\sum_{n=-\infty}^{\infty}\left|V_R\left(y-\frac{n\pi}{R}\right)\right|\\&\le\frac{\pi}{R}\sum_{n=-\infty}^{\infty}\frac{1}{R-1}\left(R^2K\left(R\left(y-\frac{n\pi}{R}\right)\right)+K\left(y-\frac{n\pi}{R}\right)\right)\\&=\frac{1}{R-1}\left(\frac{R\pi}{R}\sum_{n=-\infty}^{\infty}RK\left(R\left(y-\frac{n\pi}{R}\right)\right)+\frac{\pi}{R}\sum_{n=-\infty}^{+\infty}K\left(y-\frac{n\pi}{R}\right)\right)\\&=\frac{R+1}{R-1}. \qquad \square\end{aligned}$$

LEMMA 3. *Let $f(t) = (1/2\pi)\int_{-L}^{L} e^{i\gamma t} d\nu(\gamma)$, where ν is a bounded measure supported on $[-L, L]$. Suppose that K is an interval on the real line of the form $K = [n_1\pi/RL, n_2\pi/RL]$, where $R > 1$ and n_1 and n_2 are integers. Let M be a positive integer. Then*

$$\sup_{t\in K}|f(t)| \leq \left(\frac{R+1}{R-1}\right)\left[\max_{n_1-M\leq n\leq n_2+M}\left|f\left(\frac{n\pi}{RL}\right)\right| + \frac{8\|f\|_\infty}{\pi^2 M}\right].$$

Proof. For $t \in K$, use the interpolation result of Lemma 1 to write

$$\begin{aligned} f(t) &= \frac{\pi}{R}\sum_{n=-\infty}^{\infty} f\left(\frac{n\pi}{RL}\right) V_R\left(Lt - \frac{n\pi}{R}\right) \\ &= \frac{\pi}{R}\left(\sum_{n=-\infty}^{n_1-M-1} + \sum_{n=n_1-M}^{n_2+M} + \sum_{n=n_2+M+1}^{\infty}\right) f\left(\frac{n\pi}{RL}\right) V_R\left(Lt - \frac{n\pi}{R}\right) \\ &= S_1 + S_2 + S_3. \end{aligned}$$

Now let $\mathcal{A}$ be the set of integers n such that $n_1 - M \leq n \leq n_2 + M$, and by Lemma 2,

$$|S_2| \leq \max_{n\in\mathcal{A}}\left|f\left(\frac{n\pi}{RL}\right)\right| \frac{\pi}{R}\sum_{n=-\infty}^{\infty}\left|V_R\left(Lt - \frac{n\pi}{R}\right)\right| \leq \frac{R+1}{R-1}\max_{n\in\mathcal{A}}\left|f\left(\frac{n\pi}{RL}\right)\right|.$$

Next, consider S_3.

$$\begin{aligned} |S_3| &\leq \|f\|_\infty \frac{\pi}{R}\sum_{n=n_2+M+1}^{\infty}\left|V_R\left(Lt - \frac{n\pi}{R}\right)\right| \\ &\leq \|f\|_\infty \frac{\pi}{R}\sum_{n=n_2+M+1}^{\infty}\frac{4}{\pi(R-1)\left(Lt - n\pi/R\right)^2} \\ &\leq \|f\|_\infty \frac{\pi}{R}\sum_{n=n_2+M+1}^{\infty}\frac{4R^2}{\pi(R-1)\pi^2(n_2-n)^2} \\ &= \|f\|_\infty \frac{4R}{\pi^2(R-1)}\sum_{n=M+1}^{\infty}\frac{1}{n^2} \\ &\leq \|f\|_\infty \frac{4}{\pi^2}\left(\frac{R}{R-1}\right)\frac{1}{M}. \end{aligned}$$

We have the same estimate for $|S_1|$, so that for $t \in K$,

$$|f(t)| \leq \frac{R+1}{R-1}\max_{n\in\mathcal{A}}\left|f\left(\frac{n\pi}{RL}\right)\right| + \frac{8}{\pi^2}\left(\frac{R}{R-1}\right)\frac{\|f\|_\infty}{M}. \qquad \square$$

Remark. In the proof of the main theorem, we will apply this lemma to functions that are linear combinations of functions of the form

$$f_\omega(t) = e^{iXt} - E[e^{iXt}]$$

for random variables X taking values $X(\omega)$ in the interval $[-L, L]$. Since

$$E[e^{iXt}] = \int_{-L}^{L} e^{i\gamma t}\, dF_X(\gamma),$$

where F_X is the cumulative distribution function of X, it is clear that for each ω in the probability space $f_\omega(t)$ is the inverse Fourier transform of a measure supported on $[-L, L]$.

LEMMA 4. *Let (Ω, P) be a probability space, and suppose Z_n for $n = 1, 2, \dots, N$ are real-valued measurable functions defined on $\Omega \times \mathbf{R}$. Suppose further that for all real t,*

(1) $E[Z_n(\cdot, t)] = 0$ *for* $n = 1, 2, \dots, N$;

(2) $Z_1(\cdot, t), \dots, Z_N(\cdot, t)$ *are independent random variables*;

(3) *There exist functions $\beta_1, \dots, \beta_N$ such that for $n = 1, \dots, N$ and all real γ:*

$$E\left[e^{\gamma Z_n(\cdot,t)}\right] \le e^{\gamma^2 \beta_n(t)}.$$

Let $F(\cdot, t) = \sum_{n=1}^{N} Z_n(\cdot, t)$. Then if $\lambda > 0$ and $a > 0$, and $t_1, \dots, t_q$ are real numbers,

$$P\left[|F(\cdot, t_k)| \ge \frac{a}{\lambda} \text{ for some } k = 1, \dots, q\right] \le 2\sum_{k=1}^{q} \exp\left(\lambda^2 \sum_{n=1}^{N} \beta_n(t_k) - a\right).$$

Proof. Let $a > 0$ and $\lambda > 0$ be given. To simplify notation we will write $F(t)$ for $F(\cdot, t)$ and $Z(t)$ for $Z(\cdot, t)$. Then for each fixed real number t_k,

$$\begin{aligned} E[e^{\lambda F(t_k)}] &= E\left[\prod_{n=1}^{N} \exp(\lambda Z_n(t_k))\right] \\ &= \prod_{n=1}^{N} E[\exp(\lambda Z_n(t_k))] \le \prod_{n=1}^{N} \exp(\lambda^2 \beta_n(t_k)) \\ &= \exp\left(\lambda^2 \sum_{n=1}^{N} \beta_n(t_k)\right). \end{aligned}$$

For any nonnegative random variable X and any real number y, $y\,P[X \ge y] \le E[X]$. (This is one form of Markov's inequality.) Consequently,

$$P\left[e^{\lambda F(t_k)} \ge e^a\right] \le e^{-a} E\left[e^{\lambda F(t_k)}\right] \le \exp\left(\lambda^2 \sum_{n=1}^{N} \beta_n(t_k) - a\right),$$

or

$$P[F(t_k) \ge a/\lambda] \le \exp\left(\lambda^2 \sum_{n=1}^{N} \beta_n(t_k) - a\right).$$

To estimate $P[F(t_k) \le -a/\lambda]$, apply the same argument to the random variable $-F(t_k)$, giving

$$\begin{aligned} P[|F(t_k)| \ge a/\lambda] &= P[F(t_k) \ge a/\lambda] + P[F(t_k) \le -a/\lambda] \\ &\le 2\exp\left(\lambda^2 \sum_{n=1}^{N} \beta_n(t_k) - a\right). \end{aligned}$$

Hence

$$P[\,|F(t_k)| \ge a/\lambda \text{ for some } k \in \{1, \dots, q\}\,] \le 2\sum_{k=1}^{q} \exp\left(\lambda^2 \sum_{n=1}^{N} \beta_n(t_k) - a\right). \qquad \square$$

THEOREM. *Let $X_1, \dots, X_N$ be independent random variables taking values in the interval $[-L, L]$. For real $a_1, \dots, a_N$ and $\phi_1, \dots, \phi_N$ consider the function $F(t)$ defined by*

$$F(t) = \sum_{n=1}^{N} a_n \sin(X_n t + \phi_n),$$

and let $\mu_n(t) = E[\sin(X_n t + \phi_n)]$. Define $G(t)$ by $G(t) = F(t) - E[F(t)]$. Suppose that there exist functions $\beta_1, \dots, \beta_N$ such that for each integer $n = 1, \dots, N$ and for all real numbers γ and real values of t,

$$E[\exp(\gamma a_n(\sin(X_n t + \phi_n) - \mu_n(t)))] \le \exp(\gamma^2 \beta_n(t)).$$

Then for $a > 0, \lambda > 0, R > 1$, integer $M > 1$ and an interval $K = [n_1\pi/RL, n_2\pi/RL]$, where n_1 and n_2 are integers, we have

$$(2.3)\quad P\left[\sup_{t\in K} |G(t)| \ge \left(\frac{R+1}{R-1}\right)\left(\frac{a}{\lambda} + \frac{8\|G\|_\infty}{\pi^2 M}\right)\right] \le 2\sum_{k=1}^{q} \exp\left(\lambda^2 \sum_{n=1}^{N} \beta_n(t_k) - a\right),$$

where $t_1, \dots, t_q$ is the set of points of the form $\pi k/RL$ for $k = n_1 - M, \dots, n_2 + M$ and $q = |n_2 - n_1| + 2M + 1$.

Remark. Note that by taking $\phi_n = 0$ for each n gives $F(t)$ as the sum of $\sin X_n t$ terms, and $\phi_n = \pi/2$ for all n gives $F(t)$ as a sum of $\cos X_n t$ terms, while other values of ϕ_n give $F(t)$ as a sum of a mixture of sine and cosine terms. This observation will be used without further comment.

Proof. By definition,

$$G(t) = F(t) - E[F(t)] = \sum_{n=1}^{N} a_n(\sin(X_n t + \phi_n) - \mu_n(t)),$$

where $\mu_n(t) = E[\sin(X_n t + \phi_n)]$. Let

$$Z_n(t) = a_n(\sin(X_n t + \phi_n) - \mu_n(t)).$$

Let $a > 0, \lambda > 0, M > 1$ and an interval $K = [n_1\pi/RL, n_2\pi/RL]$ be given. Then by hypothesis, Lemma 4 applies, giving

$$(2.4)\quad P[|G(t_k)| \ge a/\lambda \text{ for some } k = 1, \dots, q] \le 2\sum_{k=1}^{q} \exp\left(\lambda^2 \sum_{n=1}^{N} \beta_n(t_k) - a\right),$$

where $t_1, \dots, t_q$ is the set of points of the form $k\pi/RL$ for $k = n_1 - M, \dots, n_2 + M$.

By Lemma 3 we have the following inclusion of events:

$$\left[\sup_{t\in K} |G(t)| < \left(\frac{R+1}{R-1}\right)\left(\frac{a}{\lambda} + \frac{8\|G\|_\infty}{\pi^2 M}\right)\right] \supset \left[|G(t_k)| < \frac{a}{\lambda} \text{ for all } k = 1, \dots, q\right].$$

Therefore, by (2.4),

$$\begin{aligned} P\left[\sup_{t\in K} |G(t)| < \left(\frac{R+1}{R-1}\right)\left(\frac{a}{\lambda} + \frac{8\|G\|_\infty}{\pi^2 M}\right)\right] &\ge P\left[|G(t_k)| < \frac{a}{\lambda} \text{ for all } k = 1, \dots, q\right] \\ &= 1 - P\left[|G(t_k)| \ge \frac{a}{\lambda} \text{ for some } k = 1, \dots, q\right] \\ &\ge 1 - 2\sum_{k=1}^{q} \exp\left(\lambda^2 \sum_{n=1}^{N} \beta_n(t_k) - a\right). \qquad \square \end{aligned}$$

We now present two useful corollaries of the main theorem. In the next section we will give some specific examples that apply these corollaries.

COROLLARY 1. *Suppose that in the main theorem* $|a_n| \leq 1$ *for* $n = 1, \ldots, N$ *and* $|\beta_n(t)| \leq B$ *for* $n = 1, \ldots, N$ *and all* t. *Then for any* $\xi_1 > 0$ *and* $\xi_2 \in (0,\ 1.6R]$,

$$P\left[\sup_{t\in K}|G(t)| \geq (\xi_1+\xi_2)\sqrt{N\log N}\right] \leq 2\left[\left(|K|L+\frac{32\rho}{\pi\xi_2}\right)\frac{R}{\pi}+3\right]N^{-(1/4B)(\xi_1/\rho)^2},$$

where $R = \sqrt{N/\log N}$ *and* $\rho = (R+1)/(R-1)$.

Proof. Let ξ_1, ξ_2, and N be given, and assume that for $n = 1, 2, \ldots, N$, $|\beta_n(t)| \leq B$ for all t. Under these hypotheses, (2.3) becomes

$$P\left[\sup_{t\in K}|G(t)| \geq \left(\frac{R+1}{R-1}\right)\left(\frac{a}{\lambda}+\frac{8\|G\|_\infty}{\pi^2 M}\right)\right] \leq 2q\exp\left(\lambda^2 NB - a\right). \tag{2.5}$$

In order to rewrite the event given in the left side of (2.5), select a and λ so that

$$\rho\frac{a}{\lambda} = \xi_1\sqrt{N\log N}. \tag{2.6}$$

Next, let $R = \sqrt{N/\log N}$, $\rho = (R+1)/(R-1)$ and $M = [16\rho R/(\pi^2\xi_2)] + 1$. The use of the greatest integer function in the definition of M and the condition that $\xi_2 \in (0,\ 1.6R]$ ensure that M is an integer larger than 1, as required in the hypotheses of the theorem. Upon substitution and using the fact that $|a_n| \leq 1$ implies that $\|G\|_\infty \leq 2N$, we have

$$\rho\frac{a}{\lambda}+\rho\frac{8\|G\|_\infty}{\pi^2 M} \leq \rho\frac{a}{\lambda}+\frac{\rho 16N\xi_2}{16\rho R} = (\xi_1+\xi_2)\sqrt{N\log N}.$$

We have, therefore, the following event inclusion:

$$\left[\sup_{t\in K}|G(t)| \geq (\xi_1+\xi_2)\sqrt{N\log N}\right] \subset \left[\sup_{t\in K}|G(t)| \geq \rho\left(\frac{a}{\lambda}+\frac{8\|G\|_\infty}{\pi^2 M}\right)\right].$$

It follows from (2.5) that

$$P\left[\sup_{t\in K}|G(t)| \geq (\xi_1+\xi_2)\sqrt{N\log N}\right] \leq 2q\exp\left(\lambda^2 NB - a\right). \tag{2.7}$$

This bound holds for all positive values of a and λ, so we will minimize it, subject to the constraint given by (2.6) that $a/\lambda = \xi_1\sqrt{N\log N}/\rho$. If we substitute in for a in (2.7) and simply consider the exponent, we must select λ to minimize the function $f(\lambda)$ given by

$$f(\lambda) = \lambda^2 NB - \lambda\frac{\xi_1}{\rho}\sqrt{N\log N}.$$

The minimum occurs at the following values for λ and a:

$$\lambda = \frac{\xi_1}{2B\rho}\sqrt{\frac{\log N}{N}} \quad\text{and}\quad a = \frac{1}{2B}\left(\frac{\xi_1}{\rho}\right)^2\log N.$$

When these values are substituted in (2.7) the resulting bound is

$$2q\exp\left(\frac{1}{4B}\left(\frac{\xi_1}{\rho}\right)^2\log N - \frac{1}{2B}\left(\frac{\xi_1}{\rho}\right)^2\log N\right) = 2qN^{-(1/4B)(\xi_1/\rho)^2}. \tag{2.8}$$

Finally, observe that $q = |n_2 - n_1| + 2M + 1$ and $|K| = |n_2 - n_1|\pi/RL$, and use the bound from (2.8) along with this value of q in (2.7) to obtain the result. □

Remark. In Examples (a)–(b) of the next section Corollary 1 is used, with a typical value of B being 0.5. In Example (c), where Corollary 2 is used, we will use a fairly modest restriction on $|K|$ to give small enough values of B to significantly improve the resulting estimates.

COROLLARY 2. *Suppose that in the main theorem we have* $|a_n| \le 1$, $|X_n| \le L$, $n_1 = 0$, *and that for some* $\theta > 0$,

$$|\beta_n(t)| \le \theta\, t^2 \quad \textit{for all } n = 1, 2, \dots, N \quad \textit{and all real } t.$$

Assume that the set $K = [0, n_2\pi/RL]$ *is selected so that* $t_q \equiv (n_2 + M)\pi/RL$ *satisfies the inequality* $\theta\, t_q^2 \le B$. *Then, with* ξ_1, ξ_2, R, ρ, *and*M *defined in the same manner as in Corollary* 1,

$$P\left[\sup_{t\in K} |G(t)| \ge (\xi_1 + \xi_2)\sqrt{N \log N}\right] \le 2\left[\left(|K|L + \frac{32\rho}{\pi\xi_2}\right)\frac{R}{\pi} + 3\right] N^{-(1/4B)(\xi_1/\rho)^2}.$$

Proof. Let ξ_1, ξ_2, and N be given, and assume that for some $\theta > 0$, $|\beta_n(t)| \le \theta t^2$ for each t, for $n = 1, \dots, N$. Then, under the remaining hypotheses of the Corollary, for each t_k of the form $k\pi/RL, k = -M, \dots, n_2 + M$, and for each $n = 1, \dots, N$ we have

$$|\beta_n(t_k)| \le \theta\, t_k^2 \le \theta\, t_q^2 \le B. \tag{2.9}$$

The inequalities given by (2.9) allow us to use (2.3) to derive the bound corresponding to (2.5) in the proof of Corollary 1. Now follow the same proof as in the rest of Corollary 1, using the same definitions of R, ρ, M, a, and λ. □

3. Examples. In this section we give several examples that apply the two corollaries of the preceding section. The main task in applying our results is to obtain good estimates for the functions $\beta_n(t)$. We provide two examples to illustrate Corollary 1, in which $\beta_n(t) \le \frac{1}{2}$ for $n = 1, 2, \dots, N$ for each t. Our third example, illustrating Corollary 2, leads to the estimate $\beta_n(t) \le \theta t^2$ for some positive θ and shows that the bound for the probability being estimated can be lowered significantly, depending upon how the set K and the random variables X_n are chosen.

In each of the examples to follow we define $Z_n(t)$ by

$$Z_n(t) = a_n(\sin(X_n t + \phi_n) - \mu_n(t)),$$

where $\mu_n(t) = E[\sin(X_n t + \phi_n)]$, and a_n is real. In any application of our results there are a number of tradeoffs possible between the various parameters involved. Specifically, these parameters are N, the number of array elements; $(\xi_1 + \xi_2)\sqrt{N \text{log} N}$, the desired level for the supremum; $|K|$, the Lebesgue measure of the set K; the location of the set K; the aperture width $2L$; the manner in which the X_n are distributed in $[-L, L]$; and the desired upper bound on the probability being estimated. We do not explore all the possibilities, but will illustrate the effects of some of the parameters. In particular, we will see the dramatic effect that changes in the desired supremum level have upon the resulting probabilities.

Before proceeding to the details of the examples to illustrate Corollary 1, we prove a proposition that has independent interest and application beyond the present setting and which leads to a value of $B = \frac{1}{2}$ in Corollary 1. Letac [10] and Vogt [17], working independently of each other, both suggested proofs for this interesting result. The proof given herein contains elements from each of these proofs. In addition, a proof can be found in Hoeffding [7].

PROPOSITION 1. *Let Y be a random variable taking values in the interval* $[-1, 1]$ *and for which* $E[Y] = \mu$. *Then*

$$E[e^{\gamma(Y-\mu)}] \leq e^{\gamma^2/2} \quad \textit{for all real values of } \gamma.$$

Proof. It suffices to prove the result for $\gamma > 0$, since the proposition is trivially true for $\gamma = 0$, and for $\gamma < 0$ the following proof can be applied to the random variable $Z = -Y$. Hence, for $\gamma > 0$ define the function $k(\gamma)$ by

$$k(\gamma) = \log(E[e^{\gamma Y}]).$$

Next, calculate $k'(\gamma)$ and $k''(\gamma)$:

$$\begin{aligned} k'(\gamma) &= \frac{E\left[Ye^{\gamma Y}\right]}{E\left[e^{\gamma Y}\right]} \qquad \text{and} \\ k''(\gamma) &= \frac{E\left[e^{\gamma Y}\right]E\left[Y^2e^{\gamma Y}\right] - \left(E\left[Ye^{\gamma Y}\right]\right)^2}{\left(E\left[e^{\gamma Y}\right]\right)^2} \\ &= \frac{E\left[Y^2e^{\gamma Y}\right]}{E\left[e^{\gamma Y}\right]} - (k'(\gamma))^2 \\ &\leq 1 - (k'(\gamma))^2 \quad (\text{since } |Y| \leq 1) \\ &\leq 1. \end{aligned}$$

Observe that the definition of k gives $k(0) = 0$ and $k'(0) = \mu$.

Now we apply the fundamental theorem of calculus to both k and k'. Since $\gamma > 0$ and $k''(t) \leq 1$ for $t > 0$ we can write

$$\begin{aligned} k'(\gamma) &= k'(0) + \int_0^\gamma k''(t)dt \leq \mu + \gamma \quad \text{and} \\ k(\gamma) &= k(0) + \int_0^\gamma k'(t)dt \leq \int_0^\gamma (\mu + t)dt = \mu\gamma + \frac{\gamma^2}{2}. \end{aligned}$$

Therefore, for $0 < \gamma < \infty$ we have, from the definition of $k(\gamma)$,

$$\begin{aligned} \log\left(E\left[e^{\gamma Y}\right]\right) &\leq \mu\gamma + \frac{\gamma^2}{2}, \quad \text{or} \\ E\left[e^{\gamma Y}\right] &\leq e^{\mu\gamma + \gamma^2/2} \\ E\left[e^{\gamma(Y-\mu)}\right] &\leq e^{\gamma^2/2}. \qquad \square \end{aligned}$$

Kahane [8, p. 67] defines a random variable X to be a subnormal variable if

$$E[e^{\gamma X}] \leq e^{\gamma^2/2} \quad \text{for all real } \gamma.$$

Hence, Propositon 1 asserts that if Y is any random variable assuming values only in the interval $[-1, 1]$, then $Y - E[Y]$ is a subnormal variable. In particular, the random variables

$$Z_n(t) = a_n(\sin(X_nt + \phi_n) - \mu_n(t))$$

are subnormal whenever $|a_n| \leq 1$.

Example. (a) Consider

$$G(t) = \sum_{n=1}^{N} a_n \left(\sin(X_n t + \phi_n) - \mu_n(t)\right) = \sum_{n=1}^{N} Z_n(t)$$

with only the assumptions that the X_n are independent, $|X_n| \leq L$, $|a_n| \leq 1$, and ϕ_n arbitrary. According to Proposition 1,

$$E[e^{\gamma Z_n(t)}] \leq e^{\gamma^2/2} \quad \text{for all } n, \gamma, \text{and } t,$$

so that for such functions $G(t)$ we can apply Corollary 1 with $B = \frac{1}{2}$.

Before proceeding to calculation of the bounds for the probabilities in Example (a), we simplify our notation and calculations. In Corollary 1 the supremum level is being compared to the quantity $(\xi_1 + \xi_2)\sqrt{N\log N}$, with the bound for the probability being a function of ξ_1 and ξ_2. Let $l = \xi_1 + \xi_2$. With l held constant and ξ_1 varying in the interval $(0, l)$, there can be significant changes in the bound, depending upon how ξ_1 and ξ_2 are chosen. This can be seen simply by noting that ξ_1, as part of the exponent in the bound, has a more profound effect than does ξ_2. In fact, a straightforward but somewhat tedious analysis of the bound for the probability in Corollary 1 shows that ξ_2 should be chosen as a small positive number, and ξ_1 chosen very close to its upper limit of l. We do not seek the value of ξ_2 that minimizes the bound, since the calculations are cumbersome. However, for our examples, which are taken for $N = 10,000$, a convenient and good choice is to select ξ_2 so that

$$\frac{32\rho}{\pi\xi_2} = 100\sqrt{\log N}, \quad \text{whereby}$$
$$\frac{\xi_2}{\rho} = \frac{0.102}{\sqrt{\log N}} \quad \text{and} \quad \frac{\xi_1}{\rho} = \frac{l}{\rho} - \frac{0.102}{\sqrt{\log N}}.$$

If $N = 10,000$, then $R = \sqrt{N/\log N} = 32.95$, $\rho = (R+1)/(R-1) = 1.063$, and $\xi_1/\rho = l/\rho - 0.0336$. With K, L, l, and the X_n still unspecified, Corollary 1 can thus be formulated as

$$P\left[\sup_{t\in K} |G(t)| \geq l\sqrt{N\log N}\right]$$
$$\leq \frac{2\left(|K|L + 100\sqrt{\log N} + 0.288\right)}{\pi\sqrt{\log N}} \times N^{1/2-(1/4B)(0.941\,l-0.0336)^2}.$$

In the example being considered we can take $B = \frac{1}{2}$ and $\sqrt{N\log N} = 303.5$. Choosing ξ_1 and ξ_2 as indicated above and a "time-bandwidth" product of $|K|L = 1000$ yields

(3.1)
$$P\left[\sup_{t\in K}\left|\sum_{n=1}^{10,000} Z_n(t)\right| \geq 303.5\,l\right] \leq \frac{2(1,000 + 303.5 + 0.288)}{3.035\,\pi} \times 10^{2-2(0.941\,l-0.0336)^2}$$

for all collections of N phase angles ϕ_n and coefficients a_n, $|a_n| \leq 1$, and N independent random variables X_n assuming values in $[-L, L]$. In particular, if each $a_n = 1$

and $\phi_n = 0$ we have a sum of $N \sin(X_n t) - \mu_n(t)$ terms, and if each $\phi_n = \pi/2$ we have a sum of $N \cos(X_n t) - \mu_n(t)$ terms. The bound is given below for $l = 1.75$, 2, and 3. As can be seen from Table 1, the probability bounds are extremely sensitive to changes in l.

TABLE 1

l	Upper bound
1.75	1.71×10^{-1}
2	4.02×10^{-3}
3	7.51×10^{-12}

Example. (b) For t real consider the complex-valued function $G(t)$ defined by

$$G(t) = \sum_{n=1}^{N} a_n \left(e^{iX_n t} - E[e^{iX_n t}]\right),$$

where for each $n, |a_n| \leq 1$ and the X_n are independent random variables for which $|X_n| \leq L$. Then, for any nonnegative number a and for each fixed $t \in K$ we have the following set inclusion in the probability space Ω:

$$\left[\omega : |G(\omega, t)| \geq a\sqrt{2}\right] \subset \left[\omega : |\Re G(\omega, t)| \geq a\right] \cup \left[\omega : |\Im G(\omega, t)| \geq a\right].$$

Consequently, we can apply the result of Example (a) to both the real and imaginary parts of $G(t)$ to bound the right-hand side of:

$$P\left[\sup_{t \in K} |G(t)| \geq a\sqrt{2}\right] \leq P\left[\sup_{t \in K} |\Re G(t)| \geq a\right] + P\left[\sup_{t \in K} |\Im G(t)| \geq a\right].$$

Before proceeding to our third example, we state and prove two propositions that are used to specify the functions $\beta_n(t)$ that are needed in order to use Corollary 2.

PROPOSITION 2. *If* $\alpha \geq 0.5575$, *then*

$$e^x - x \leq e^{\alpha x^2} \quad \textit{for all real numbers } x. \tag{3.2}$$

Proof. For real x and y let

$$z(x, y) = e^{yx^2} - e^x + x.$$

Note that for a fixed x, $z(x, y)$ is an increasing function of y. Hence $z(x, y)$ is positive for all points above the curve $z(x, y) = 0$. That is,

$$\{(x, y) \,|\, z(x, y) \geq 0\} = \{(x, \xi) \,|\, \xi \geq y, \text{ where } z(x, y) = 0\}.$$

Therefore, the inequality (3.2) is valid if and only if the horizontal line $y = \alpha$ lies above the curve $z(x, y) = 0$. Thus, the best lower bound for α is y_0, where (x_0, y_0) is a global maximum for the curve $z(x, y) = 0$, which can be written as

$$y = \frac{1}{x^2} \log(e^x - x).$$

An analysis of this curve shows it to have only one critical point, which occurs at the maximum. Using Newton's method to solve $y' = 0$ gives the maximum at $x_0 = 0.6400\ldots$ and $y_0 = 0.5574\ldots$. □

PROPOSITION 3. *Let X be a random variable assuming values only in the interval $[-L, L]$, and suppose that for each real t the random variable $Z(t) = \sin(Xt) - E[\sin(Xt)]$. Suppose that for some constant $w > 0$ (independent of t and j) the following condition holds for every real t:*

$$|E[Z^j(t)]| \leq |wt|^j \quad \textit{for } j = 1, 2, 3, \ldots .$$

Then for each real t we have the following bound:

$$E\left[e^{\gamma Z(t)}\right] \leq e^{\gamma^2 \alpha (wt)^2} \quad \textit{for all real } \gamma,$$

where α is given by Proposition 2.

Proof. The conditions on Z and its moments allow us to write

$$\begin{aligned} E\left[e^{\gamma Z(t)}\right] &= E\left[1 + \gamma Z(t) + \sum_{j=2}^{\infty} \frac{(\gamma Z(t))^j}{j!}\right] \\ &\leq 1 + \sum_{j=2}^{\infty} \frac{|\gamma|^j \, |E[Z^j(t)]\,|}{j!} \\ &\leq 1 + \sum_{j=2}^{\infty} \frac{|\gamma|^j \, |wt|^j}{j!} = e^{|\gamma w t|} - |\gamma w t|. \end{aligned}$$

By Proposition 2, the last expression is bounded above by $\exp\left(\alpha \gamma^2 (wt)^2\right)$. □

Remark. In Proposition 3 we always have $E\left[e^{\gamma Z(t)}\right] \leq e^{\gamma^2/2}$, by Proposition 1. By suitably restricting the set K from which the t values are chosen so that, for some B, $\alpha(wt)^2 \leq B < \frac{1}{2}$ it is possible to improve upon the estimates given by Corollary 1. In this way we exploit the bounds given by Corollary 2 in our concluding example.

Example. (c) Our final example illustrates an application of Corollary 2. For even, positive integers N define N random variables X_n by

$$X_n = Y_n + nw \text{ for integers } n = \frac{-N}{2}, \ldots, \frac{N}{2} - 1,$$

where the Y_n are uniformly distributed on the interval $(0, w)$. Now let $w = 2L/N$, so that the interval $[-L, L]$ is divided into N bins of width w with X_n uniformly distributed on the nth bin. Let

$$Z_n(t) = \sin X_n t - \mu_n(t).$$

By the integral mean value theorem, for each real t and each integer n, we have

$$\mu_n(t) = \frac{1}{w} \int_0^w \sin(nw + y)t \, dy = \sin \xi$$

for some $\xi \in (\, nwt, (n+1)wt\,)$. Hence

$$E[Z_n^j(t)] = \frac{1}{w} \int_0^w (\sin(nw + y)t - \sin \xi)^j \, dy.$$

By the derivative mean value theorem

$$|\sin(nw+y)t-\sin\xi| \leq |wt|, \quad \text{so that}$$

$$\left|E[Z_n^j(t)]\right| \leq |wt|^j.$$

According to Proposition 3 we have

$$E\left[e^{\gamma Z_n(t)}\right] \leq \exp\left(\alpha\gamma^2|wt|^2\right),$$

which gives $\beta_n(t) = \alpha w^2t^2$. Thus in Corollary 2, $\theta = \alpha w^2 = \alpha(2L/N)^2$ for this example. We are interested in sets K of the form $K = [0, n_2\pi/RL]$, where the integer n_2, which essentially determines the extent of the set K, is selected so that $t_q = (n_2+M)\pi/RL$. The restriction in Corollary 2 that $\theta t_q^2 \leq B$ becomes

$$\alpha(2L/N)^2\, t_q^2 \leq B, \quad \text{or} \quad t_q \leq \frac{N\sqrt{B}}{2L\sqrt{\alpha}}.$$

In terms of n_2, with $R = \sqrt{N/\log N}$, this gives

$$n_2+M \leq \frac{N\sqrt{B}}{2L\sqrt{\alpha}} \cdot \frac{L\sqrt{N}}{\pi\sqrt{\log N}} \leq \frac{0.214\sqrt{B}\,N^{3/2}}{\sqrt{\log N}}.$$

To compare Examples (a) and (c) we use the same substitutions for N, M, ξ_1, and ξ_2 as in Example (a) and Corollary 1. Therefore,

$$N = 10,000, \quad \frac{\rho}{\xi_2} = \frac{\sqrt{\log N}}{0.102}, \quad R = 32.95, \quad \text{and} \quad M = \left[\frac{16\,R\rho}{\pi^2\xi_2}\right]+1 = 1592,$$

giving

$$n_2 \leq 0.0706\sqrt{B}\,10^6 - M \leq 7.06\sqrt{B}\,10^4 - 1,592.$$

To illustrate, let $B = \frac{1}{4}$. Then any $n_2 \leq 33,708$ gives a set K satisfying the hypotheses of Corollary 2. In Example (c), let $n_2 = 32,000$, so that $K = [0, 3.2\times 10^4\,\frac{\pi}{RL}]$. Thus, $|K|L = 32,000\frac{\pi}{R} = 3051$, and the estimate in Corollary 2 corresponding to (3.1) is

$$P\left[\sup_{t\in K}\left|\sum_n \sin(X_nt)-\mu_n(t)\right| \geq 303.5\,l\right] \leq \frac{2(3051+303.5+0.288)}{3.035\pi} \times 10^{2-4(0.941l-0.0336)^2}.$$

We now tabulate the bounds for the two Corollaries using the above set K, but different $B's$. (See Table 2.) The estimate for Corollary 1 uses $B = \frac{1}{2}$ and $|K|L = 3051$ to obtain the analogue of (3.1). The estimate for Corollary 2 uses $B = \frac{1}{4}$. The Corollary 1 estimate applies to any collection of X_n distributed independently on $[-L, L]$, while the Corollary 2 estimate is valid for X_n distributed according to the bin approach as described in Example (c).

TABLE 2

l	Cor. 1 Upper bound	Cor. 2 Upper bound
1.50	1.13×10^{-0}	1.80×10^{-3}
1.75	4.39×10^{-1}	2.75×10^{-6}
2.00	1.03×10^{-2}	1.52×10^{-9}

In the case of the Corollary 1 upper bound, the value of $l = 1.5$ gives a bound larger than 1, so a larger value of l needs to be chosen. A comparison of the bounds derived from the two Corollaries shows that in the bin example for t near zero, the bounds for the probabilities of deviation from the mean are lowered significantly. An explanation for this lies in the fact that the bin approach provides more control in the manner in which the X_n are distributed, resulting in less chance of a large deviation when t is small. Plots of the variance of $G(t)$, as a function of t, also show that when t is near zero the variance under the bin approach is considerably smaller than, for example, when the X_n are distributed independently and uniformly across the entire aperture $[-L, L]$, though for larger t the two variances approach the same limiting value of $N/2$ as t increases. Further details and plots of these phenomena can be found in Hendricks [6].

REFERENCES

[1] L. BREIMAN, *Probability*, Addison-Wesley, Reading, MA, 1968.

[2] K. L. CHUNG, *A Course in Probability Theory, Second Edition*, Academic Press, New York, 1974.

[3] M. B. DONVITO AND S. A. KASSAM, *Characterization of the random array peak sidelobe*, IEEE Trans. Antennas and Propagation, AP-27 (1979), pp. 379–385.

[4] W. FELLER, *An Introduction to Probability Theory and Its Applications, Vol.* 2, John Wiley, New York, 1966.

[5] S. HAYKIN, ED., *Array Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1985.

[6] W. J. HENDRICKS, *The totally random versus the bin approach for random arrays*, IEEE Trans. Antennas and Propagation, AP-39 (1991), pp. 1757–1762.

[7] W. HOEFFDING, *Probability inequalities for sums of bounded random variables*, Amer. Statist. Assoc. J., 58 (1963), pp. 13-30.

[8] J. P. KAHANE, *Some Random Series of Functions, Second Edition*, Cambridge University Press, Cambridge, England, 1985.

[9] Y. KATZNELSON, *An Introduction to Harmonic Analysis*, John Wiley, New York, 1968.

[10] G. LETAC, personal communication, Dèp. de Mathèmatique, Univ. Paul Sabatier, Toulouse, France, 1991.

[11] Y. T. LO, *A mathematical theory of antenna arrays with randomly spaced elements*, IEEE Trans. Antennas and Propagation, AP-12 (1964), pp. 257–268.

[12] ———, *A probabilistic approach to the problem of large antenna arrays*, Radio Science J. Res. NBS/USNC-URSI, 68D (1964), pp. 1011–1019.

[13] Y. T. LO AND V. D. AGRAWAL, *Distribution of sidelobe levels in random arrays*, Proc. IEEE, 57 (1969), pp. 1764–1765.

[14] R. SALEM AND A. ZYGMUND, *Some properties of trigonometric series whose terms have random signs*, Acta Math., 91 (1954), pp. 245–301.

[15] B. D. STEINBERG, *Principles of Aperture & Array System Design*, John Wiley, New York, 1976.

[16] ———, *The peak sidelobe of the phased array having randomly located elements*, IEEE Trans. Antennas and Propagation, AP-20 (1972), pp. 129–136.

[17] A. VOGT, personal communication, Dept. of Mathematics, Georgetown University, Washington, DC, 1991.

[18] A. ZYGMUND, *Trigonometric Series*, Cambridge University Press, Cambridge, England, 1959.

SELBERG INTEGRALS AND HYPERGEOMETRIC FUNCTIONS ASSOCIATED WITH JACK POLYNOMIALS*

JYOICHI KANEKO†

Abstract. A new class of hypergeometric functions of several variables is introduced by using Jack polynomials and a multivariate generalization of Aomoto's result iş given [*SIAM J. Math. Anal.*, 18 (1987), pp. 545-549].

Key words. Selberg integrals, Jack polynomials, hypergeometric functions

AMS subject classifications. 33A15, 33A35, 33A65

1. Introduction. Let $D_{\lambda_1,\lambda_2,\lambda}(x_1,\ldots,x_n)$ be the Selberg density:

$$\prod_{i=1}^{n} x_i^{\lambda_1}(1-x_i)^{\lambda_2} \prod_{1\le i<j\le n} |x_i-x_j|^{\lambda}, \quad \lambda_1,\lambda_2>-1, \quad \lambda>0,$$

and put

$$S_{n,m}(\lambda_1,\lambda_2,\lambda,\mu\colon t_1,\ldots,t_m) \quad (S_{n,m}(t) \text{ for short})$$
$$=\int_{[0,1]^n} \prod_{\substack{1\le i\le n\\1\le k\le m}} (x_i-t_k)^{\mu} D_{\lambda_1,\lambda_2,\lambda}(x_1,\ldots,x_n)\,dx_1\cdots dx_n.$$

We assume for simplicity that t_i, $1\le i\le m$, lies in $\mathbb{C}-[0,1]$ if μ is not a nonnegative integer.

The celebrated formula of Selberg [Se] is

$$(1)\quad S_{n,0}(\lambda_1,\lambda_2,\lambda)=\prod_{i=1}^{n}\frac{\Gamma(i(\lambda/2)+1)\Gamma(\lambda_1+1+(i-1)(\lambda/2))\Gamma(\lambda_2+1+(i-1)(\lambda/2))}{\Gamma((\lambda/2+1)\Gamma(\lambda_1+\lambda_2+2+(n+i-2)(\lambda/2))}.$$

Selberg's evaluation had come to light around 1980 from a dormancy of nearly forty years. Since then this type of integral with Selberg density and its generalizations have been studied intensively by many authors, e.g., [Ao1], [Ao2], [Ao3], [As], [Kad], and [Ma2]. Among them we must mention a result of Aomoto [Ao1]:

$$S_{n,1}(\lambda_1,\lambda_2,\lambda,1;t)=S_{n,0}(\lambda_1,\lambda_2,\lambda)\frac{n!}{\prod_{i=1}^{n}(\alpha+\beta+n+i)}P_n^{(\alpha,\beta)}(1-2t),$$

where $P_n^{(\alpha,\beta)}(x)$ denotes the Jacobi polynomial of degree n and $\alpha=-1+2(\lambda_1+1)/\lambda$, $\beta=-1+2(\lambda_2+1)/\lambda$. The purpose of the present paper is to give a multivariate generalization of this result. Namely, we show that $S_{n,m}(t)$ with $\mu=1$ or $\mu=-\lambda/2$ can be expressed by the hypergeometric function constructed from Jack polynomials which will be defined in § 3.2. (For the precise form of the expression see Theorem 5 in § 6.1.) The rough idea of proof is as follows. First we calculate a holonomic system for $S_{n,m}(t)$ (Theorem 1). Also, it can be verified that the hypergeometric function ${}_2F_1^{(\alpha)}(a,b;c;t_1,\ldots,t_m)$ referred to above satisfies a holonomic system of the same form (Theorem 4 in § 4.3). Then we can adjust the parameters α, a, b, c to make these two systems identical. Therefore, the desired expression follows from the uniqueness property of the solution (Theorem 2 in § 4.1). As a by-product we give a new proof of the integration formula of Jack polynomials conjectured by Macdonald [Ma3] and

* Received by the editors March 13, 1991; accepted for publication (in revised form) October 19, 1992.

† Department of Mathematics, College of General Education, Kyushu University, Fukuoka 810, Japan.

proved by Kadell [Kad]. In § 7 we show that $S_{n.m}(t)$ with $\mu=1$ is a special case of generalized Jacobi polynomials defined and studied by Koornwinder [Ko], Vretare [V], and Debiard [D].

While preparing this paper, the author was notified by A. Korányi that he had introduced the same hypergeometric functions ${}_2F_1^{(\alpha)}(a, b; c; t_1, \dots, t_m)$ in [Kor] and that his student Z. Yan also had proved our Theorem 4 by a similar method in his thesis (see [Y]).

2. Holonomic system for $S_{n,m}(t)$.

2.1. Holonomic system. We denote by Φ the function $D_{\lambda_1,\lambda_2,\lambda}(x)\prod_{1\le i\le n, 1\le k\le m}(x_i-t_k)^{\mu}$, and by ω the logarithmic 1-form $d\log\Phi$. Let ∇_ω be the covariant differentiation defined by $\nabla_\omega\varphi = d\varphi+\omega\wedge\varphi$ for an $(n-1)$-form φ. Since $d(\Phi\varphi)=\Phi\nabla_\omega\varphi$, the Stokes formula gives

$$\int_{[0,1]^n}\Phi\nabla_\omega\varphi=0 \tag{2}$$

as long as the left-hand side exists. Let us denote by $^*dx_i$ the $(n-1)$-form $(-1)^{i-1}dx_1\wedge\cdots\wedge dx_{i-1}\wedge dx_{i+1}\wedge\cdots\wedge dx_n$, and put

$$\varphi_0=\sum_{i=1}^n {}^*dx_i,$$

$$\varphi_1=\sum_{i=1}^n x_i\,{}^*dx_i,$$

$$\psi_k=\sum_{i=1}^n (x_i-t_k)^{-1}\,{}^*dx_i,\qquad 1\le k\le m.$$

The covariant differentiation of these forms are

$$\nabla_\omega\varphi_0=\left[\lambda_1\sum_{i=1}^n x_1^{-1}-\lambda_2\sum_{i=1}^n(1-x_i)^{-1}+\mu\sum_{\substack{1\le i\le n\\1\le k\le m}}(x_i-t_k)^{-1}\right]\theta, \tag{3}$$

$$\nabla_\omega\varphi_1=\left[n\left(1+\lambda_1+\lambda_2+m\mu+\frac{n-1}{2}\lambda\right)-\lambda_2\sum_{i=1}^n(1-x_i)^{-1}+\mu\sum_{\substack{1\le i\le n\\1\le k\le m}}\frac{t_k}{x_i-t_j}\right]\theta, \tag{4}$$

$$\begin{aligned}\nabla_\omega\psi_k=\Bigg[&(\mu-1)\sum_{i=1}^n(x_i-t_k)^{-2}-\lambda\sum_{1\le i<j\le n}((x_i-t_k)(x_j-t_k))^{-1}+\lambda_1 t_k^{-1}\\ &\cdot\left(\sum_{i=1}^n(x_i-t_k)^{-1}-\sum_{i=1}^n x_i^{-1}\right)-\lambda_2(1-t_k)^{-1}\left(\sum_{i=1}^n(1-x_i)^{-1}+\sum_{i=1}^n(x_i-t_k)^{-1}\right)\\ &+\mu\sum_{\substack{l=1\\l\ne k}}^m(t_k-t_l)^{-1}\left(\sum_{i=1}^n(x_i-t_k)^{-1}-\sum_{i=1}^n(x_i-t_l)^{-1}\right)\Bigg]\theta,\end{aligned} \tag{5}$$

where θ denotes the volume n-form: $\theta=dx_1\wedge\cdots\wedge dx_n$. For n-forms ξ, η, we write $\xi\sim\eta$ if $\xi-\eta=\nabla_\omega\varphi$ for some $(n-1)$-form φ. It follows from (3) and (4) that

$$\left[\lambda_1\sum_{i=1}^n x_i^{-1}\right]\theta\sim\left[n\left(1+\lambda_1+\lambda_2+m\mu+\frac{n-1}{2}\lambda\right)-\mu\sum_{\substack{1\le i\le n\\1\le k\le m}}\frac{1-t_k}{x_i-t_k}\right]\theta,$$

$$\left[\lambda_2\sum_{i=1}^n(1-x_i)^{-1}\right]\theta\sim\left[n\left(1+\lambda_1+\lambda_2+m\mu+\frac{n-1}{2}\lambda\right)+\mu\sum_{\substack{1\le i\le n\\1\le k\le m}}\frac{t_k}{x_i-t_k}\right]\theta.$$

Substituting these into (5), we obtain

$$
\begin{aligned}
\nabla_\omega \psi_k \sim \Bigg[& (\mu-1)\sum_{i=1}^{n}(x_i-t_k)^{-2}-\lambda\sum_{1\le i<j\le n}((x_i-t_k)(x_j-t_k))^{-1} \\
& +(\lambda_1 t_k^{-1}-\lambda_2(1-t_k)^{-1})\left(\sum_{i=1}^{n}(x_i-t_k)^{-1}\right) \\
& -t_k^{-1}\left(n\left(1+\lambda_1+\lambda_2+m\mu+\frac{n-1}{2}\lambda\right)-\mu\sum_{\substack{1\le i\le n\\ 1\le l\le m}}\frac{1-t_l}{x_i-t_l}\right) \\
& -(1-t_k)^{-1}\left(n\left(1+\lambda_1+\lambda_2+m\mu+\frac{n-1}{2}\lambda\right)+\mu\sum_{\substack{1\le i\le n\\ 1\le l\le m}}\frac{t_l}{x_i-t_l}\right) \\
& +\mu\sum_{\substack{l=1\\ l\ne k}}^{m}(t_k-t_l)^{-1}\left(\sum_{i=1}^{n}(x_i-t_k)^{-1}-\sum_{i=1}^{n}(x_i-t_l)^{-1}\right)\Bigg]\theta.
\end{aligned} \tag{6}
$$

On the other hand, one can easily show that

$$
\frac{\partial S_{n,m}(t)}{\partial t_k}=-\mu\int_{[0,1]^n}\Phi\left[\sum_{i=1}^{n}(x_i-t_k)^{-1}\right]\theta, \tag{7}
$$

$$
\begin{aligned}
\frac{\partial^2 S_{n,m}(t)}{\partial t_k^2}=\int_{[01]^n}\Phi\Bigg[& (\mu^2-\mu)\sum_{i=1}^{n}(x_i-t_k)^{-2} \\
& +2\mu^2\sum_{1\le i<j\le n}((x_i-t_k)(x_j-t_k))^{-1}\Bigg]\theta.
\end{aligned} \tag{8}
$$

Suppose now that the ratio $(\mu^2-\mu)/2\mu^2$ equals $(\mu-1)/(-\lambda)$, i.e., $\mu=1$ or $\mu=-\lambda/2$. Then by (8) the first two sums of the right-hand side of (6) add up to a constant multiple of $\partial^2 S_{n,m}(t)/\partial t_k^2$. Hence, by virtue of (7) and (8), taking ψ_k for φ of (2) yields a partial differential equation of $S_{n,m}(t)$ for each k. Moreover, its principal part contains only $\partial^2 S_{n,m}(t)/\partial t_k^2$. Thus we have proved the following.

THEOREM 1. *Assume* $\mu=1$ *or* $\mu=-\lambda/2$. *Then* $S_{n,m}(\lambda_1,\lambda_2,\lambda,\mu;\, t)$ *satisfies the following holonomic system*:

$$
\begin{aligned}
& t_i(1-t_i)\frac{\partial^2 F}{\partial t_i^2}+\left\{c-\frac{1}{\alpha}(m-1)-\left(a+b+1-\frac{1}{\alpha}(m-1)\right)t_i\right\}\frac{\partial F}{\partial t_i}-abF \\
& +\frac{1}{\alpha}\left\{\sum_{\substack{j=1\\ j\ne i}}^{m}\frac{t_i(1-t_i)}{t_i-t_j}\frac{\partial F}{\partial t_i}-\sum_{\substack{j=1\\ j\ne i}}^{m}\frac{t_j(1-t_j)}{t_i-t_j}\frac{\partial F}{\partial t_j}\right\}=0, \qquad i=1,\dots,m,
\end{aligned} \tag{9}
$$

where $\alpha=\lambda/2$, $a=-n$, $b=(2/\lambda)(\lambda_1+\lambda_2+m+1)+n+1$, $c=(2/\lambda)(\lambda_1+m)$ *if* $\mu=1$, *and* $\alpha=2/\lambda$, $a=(\lambda/2)n$, $b=-(\lambda_1+\lambda_2+1)+\lambda/2(m-n+1)$, $c=-\lambda_1+(\lambda/2)m$ *if* $\mu=-\lambda/2$.

Remark. The rank of the holonomic system (9) is 2^m. This and related matters will be treated in a forthcoming paper.

2.2. Cases of $m=1,2$. In case $m=1$, (9) is nothing but the Gauss hypergeometric equation. Hence if $\mu=1$ the result of Aomoto [Ao1] mentioned above follows at once from (9). In case $m=2$, (9) is related with the holonomic system (F_4) of Appell's hypergeometric function F_4 [A-K] in the following way. The system (F_4) is

$$x(1-x)\frac{\partial^2 F}{\partial x^2}-y^2\frac{\partial^2 F}{\partial y^2}-2xy\frac{\partial^2 F}{\partial x\,\partial y}+(c_1-(a+b+1)x)\frac{\partial F}{\partial x}$$

$$-(a+b+1)y\frac{\partial F}{\partial y}-abF=0,$$

$$y(1-y)\frac{\partial^2 F}{\partial y^2}-x^2\frac{\partial^2 F}{\partial x^2}-2xy\frac{\partial^2 F}{\partial x\,\partial y}+(c_2-(a+b+1)y)\frac{\partial F}{\partial y}$$

$$-(a+b+1)x\frac{\partial F}{\partial x}-abF=0.$$

Let $\phi:\mathbb{C}^2\to\mathbb{C}^2$ be defined by $\phi(t_1,t_2)=(t_1t_2,(1-t_1)(1-t_2))=(x,y)$. Then $\phi^*(F_4)$, the pull back of the system (F_4) by ϕ, is [Kat]

$$(t_1-t_2)\left\{t_1(1-t_1)\frac{\partial^2 F}{\partial t_1^2}+(c_1-(a+b+1)t_1)\frac{\partial F}{\partial t_1}-abF\right\}+\varepsilon(t_2-1)\left(t_1\frac{\partial F}{\partial t_1}-t_2\frac{\partial F}{\partial t_2}\right)=0,$$

$$(t_2-t_1)\left\{t_2(1-t_2)\frac{\partial^2 F}{\partial t_2^2}+(c_1-(a+b+1)t_2)\frac{\partial F}{\partial t_2}-abF\right\}+\varepsilon(t_1-1)\left(t_2\frac{\partial F}{\partial t_2}-t_1\frac{\partial F}{\partial t_1}\right)=0,$$

where $\varepsilon=c_1+c_2-a-b-1$. Putting $c_1=c-(1/\alpha)$ and $c_2=a+b-c+1$ we see that $\phi^*(F_4)$ reduces to the system (9). It is known that if c_1 and c_2 are not integers, then (F_4) has the following four linearly independent solutions near the origin [A-K, p. 52]:

$$F_4(a,b,c_1,c_2;x,y)=\sum_{m,n=0}^{\infty}\frac{(a)_{m+n}(b)_{m+n}}{(c_1)_m(c_2)_n}\frac{x^my^n}{m!\,n!},$$

$$\begin{aligned}&x^{1-c_1}F_4(a-c_1+1,b-c_1+1,2-c_1,c_2;x,y),\\ (10)\qquad &y^{1-c_2}F_4(a-c_2+1,b-c_2+1,c_1,2-c_2;x,y),\\ &x^{1-c_1}y^{1-c_2}F_4(a-c_1-c_2+2,b-c_1-c_2+2,2-c_1,2-c_2;x,y),\end{aligned}$$

where $(a)_n=\Gamma(a+n)/\Gamma(a)$. Hence when $\mu=1$ we get

$$S_{n,2}(\lambda_1,\lambda_2,\lambda,1;t_1,t_2)=(-1)^n\cdot S_{n,0}(\lambda_1+1,\lambda_2+1,\lambda)$$

$$\cdot F_4(a,b,c_1,c_2;t_1t_2,(1-t_1)(1-t_2)),$$

where $a=-n$, $b=(2/\lambda)(\lambda_1+\lambda_2+3)+n-1$, $c_1=(2/\lambda)(\lambda_1+1)$, $c_2=(2/\lambda)(\lambda_2+1)$.

The case of $\mu=-\lambda/2$ will be treated in § 6.2.

3. Hypergeometric function.

3.1. Jack symmetric functions. We recall here the basic properties of Jack symmetric functions. Our reference is the fundamental [St].

Let $\kappa=(\kappa_1,\kappa_2,\ldots)$ be a partition and $\kappa'=(\kappa_1',\kappa_2',\ldots)$ be the conjugate partition. The number κ_1' of parts of κ is denoted by $l(\kappa)$ and called the length of κ. We write $|\kappa|=d$ if $\kappa_1+\kappa_2+\cdots=d$. If ν is another partition, then write $\nu\leq\kappa$ if $|\nu|=|\kappa|$ and $\nu_1+\nu_2+\cdots+\nu_i\leq\kappa_1+\kappa_2+\cdots+\kappa_i$ for all i. If κ has $m_i=m_i(\kappa)$ parts equal to i, then write

$$z_\kappa=(1^{m_1}2^{m_2}\cdots)m_1!\,m_2!\cdots.$$

Let $t=(t_1, t_2, \ldots)$ be an infinite set of indeterminates. We define the *minomial symmetric function* $m_\kappa = m_\kappa(t)$ indexed by partition $\kappa = (\kappa_1, \kappa_2, \ldots)$ by

$$m_\kappa = \sum t_1^{\kappa_1} t_2^{\kappa_2} \ldots,$$

where the summation sign indicates that we are to form all distinct monomials in t_i with exponent $\kappa_1, \kappa_2, \ldots$. The *power sum symmetric function* p_κ is defined by

$$p_\kappa = p_{\kappa_1} p_{\kappa_2}, \ldots, \qquad p_d = \sum t_i^d = m_{(d)},$$

where (d) denotes the partition $(d, 0, 0, \ldots)$.

Let α be a parameter, and let $\mathbb{Q}(\alpha)$ denote the field of all rational functions of α with rational coefficients. We define a scalar product $\langle\,,\rangle$ on the vector space of all symmetric functions of t of bounded degree over the field $\mathbb{Q}(\alpha)$ by the condition

$$\langle p_\kappa, p_\nu \rangle = \delta_{\kappa\nu} z_\kappa \alpha^{l(\kappa)},$$

where $\delta_{\kappa\nu} = 0$ if $\kappa \neq \nu$ and $\delta_{\kappa\kappa} = 1$. The *Jack symmetric function* $J_\kappa = J_\kappa^{(\alpha)}(t)$ are defined to be the unique symmetric function satisfying the following three conditions.

(P1) (orthogonality). $\langle J_\kappa, J_\nu \rangle = 0$ if $\kappa \neq \nu$.

(P2) (triangularity). If we write J_κ in terms of the monomial symmetric functions $J_\kappa = \sum_\nu v_{\kappa\nu}(\alpha) m_\nu$, then $v_{\kappa\nu}(\alpha) = 0$ unless $\nu \leq \kappa$.

(P3) (normalization). If $|\kappa| = d$, then the coefficient $v_{\kappa,1}d$ of $t_1 t_2 \cdots t_d$ in J_κ is equal to $d!$.

Setting all but finitely many variables equal to 0 (say $t_{m+1} = t_{m+2} = \cdots = 0$) in J_κ, we obtain a polynomial $J_\kappa^{(\alpha)}(t_1, \ldots, t_m)$ which we call the *Jack polynomial.* Note that the Jack polynomials $J_\kappa(t_1, \ldots, t_r)$ vanish for $r < l(\kappa)$ and are linearly independent otherwise [St, Prop. 2.5]. We will write J_n for $J_{(n)}$. Clearly we have $J_1 = t_1 + t_2 + \cdots$. Define the differential operator $D(\alpha)$ by

$$D(\alpha) = \frac{\alpha}{2} \sum_{i=1}^{m} t_i^2 \frac{\partial^2}{\partial t_i^2} + \sum_{\substack{1 \leq i,j \leq m \\ i \neq j}} \frac{t_i^2}{t_i - t_j} \frac{\partial}{\partial t_i}. \tag{11}$$

Then the Jack polynomials $J_\kappa^{(\alpha)}(t_1, \ldots, t_m)$ with $l(\kappa) \leq m$ are eigenfunctions of $D(\alpha)$ [St, Thm. 3.1]:

$$D(\alpha) J_\kappa^{(\alpha)}(t_1, \ldots, t_m) = e_\kappa(\alpha) J_\kappa^{(\alpha)}(t_1, \ldots, t_m), \tag{12}$$

where the eigenvalue $e_\kappa(\alpha)$ is given by

$$e_\kappa(\alpha) = \alpha \sum_{i=1}^{m} \kappa_i \frac{(\kappa_i - 1)}{2} - \sum_{i=1}^{m} (i-1)\kappa_i + (m-1)|\kappa|. \tag{13}$$

By [St, Prop. 2.3] we have

$$(t_1 + \cdots + t_m)^d = \alpha^d d! \sum_{|\kappa| = d} J_\kappa j_\kappa^{-1}, \tag{14}$$

where $j_\kappa = \langle J_\kappa, J_\kappa \rangle$. The value of j_κ is given by [St, Thm. 5.8]

$$\begin{aligned} j_\kappa &= \prod_{s \in \kappa} h_*^\kappa(s) h_\kappa^*(s), \\ h_\kappa^*(s) &= \kappa_j' - i + \alpha(\kappa_i - j + 1), \\ h_*^\kappa(s) &= \kappa_j' - i + 1 + \alpha(\kappa_i - j), \end{aligned} \tag{15}$$

where a partition κ is identified with its diagram: $\kappa = \{s=(i,j)\colon 1\le i\le l(\kappa), 1\le j\le \kappa_i\}$. We change the normalization by defining

$$C_\kappa^{(\alpha)}(t_1,\ldots,t_m)=\alpha^{|\kappa|}|\kappa|!J_\kappa^{(\alpha)}j_\kappa^{-1} \tag{16}$$

to get

$$(t_1+\cdots+t_m)^d=\sum_{|\kappa|=d}C_\kappa^{(\alpha)}(t). \tag{17}$$

We will also need the explicit formula of $C_\kappa^{(\alpha)}(t)$ evaluated at $t_1=t_2=\cdots=t_r=1$, $t_{r+1}=t_{r+2}=\cdots=t_m=0$, which we denote by $C_\kappa^{(\alpha)}(1^r)$ [St, Thm. 5.4]:

$$C_\kappa^{(\alpha)}(1^r)=\alpha^{|\kappa|}|\kappa|!j_\kappa^{-1}\prod_{(i,j)\in\kappa}(r-(i-1)+\alpha(j-1)). \tag{18}$$

3.2. Hypergeometric function. We assume from now on that $\alpha>0$. We define a generalized factorial function $[a]_\kappa^{(\alpha)}$ by

$$[a]_\kappa^{(\alpha)}=\prod_{i=1}^{l(\kappa)}\left(a-\frac{1}{\alpha}(i-1)\right)_{\kappa_i}. \tag{19}$$

Definition. Let $a_1,\ldots,a_p,b_1,\ldots,b_q$ be complex numbers such that $b_j-(1/\alpha)\times(i-1)$, $1\le j\le q$, $1\le i\le m$, are neither negative integers nor zero. The *hypergeometric function* ${}_pF_q^{(\alpha)}(a_1,\ldots,a_p;\,b_1,\ldots,b_q;\,t_1,\ldots,t_m)$ is defined by the series

$$\begin{aligned}&{}_pF_q^{(\alpha)}(a_1,\ldots,a_p;\,b_1,\ldots,b_q;\,t)\\&=\sum_{d=0}^{\infty}\sum_{|\kappa|=d}\frac{[a_1]_\kappa^{(\alpha)}\cdots[a_p]_\kappa^{(\alpha)}}{[b_1]_\kappa^{(\alpha)}\cdots[b_q]_\kappa^{(\alpha)}}\frac{C_\kappa^{(\alpha)}(t)}{d!}.\end{aligned} \tag{20}$$

Remark. The hypergeometric function ${}_pF_q^{(\alpha)}$ for $\alpha=2$ was first introduced by Herz [H] by means of Laplace transforms. The zonal polynomial expansion as in (20) was found by Constantine [C].

Next we consider the convergence of the series (20). It is necessary to estimate $C_\kappa^{(\alpha)}(t)$. For this we content ourselves with the following lemma.

Lemma 1. *Let* $\|t\|=\max\{|t_1|,\ldots,|t_m|\}$ *and* $\beta=\max\{1,1/\alpha\}$. *Then*

$$|C_\kappa^{(\alpha)}(t)|\le C\cdot\alpha^d\cdot(\beta m)^{(3/2)d}\cdot d^{m/2}\|t\|^d,\qquad d=|\kappa|, \tag{21}$$

where C denotes a constant depending only on m.

Proof. Write

$$J_\kappa^{(\alpha)}(t)=\sum_{|\nu|=d}a_\nu p_\nu,\qquad d=|\kappa|,$$

so that

$$j_\kappa=\sum_{|\nu|=d}a_\nu^2 z_\nu\alpha^{l(\nu)}.$$

By Cauchy's inequality we have

$$|J_\kappa^{(\alpha)}(t)|\le j_\kappa^{1/2}\left\{\sum_{|\nu|=d}\frac{p_\nu^2}{z_\nu\alpha^{l(\nu)}}\right\}^{1/2}. \tag{22}$$

It follows from (15) that

$$j_\kappa\ge\beta^{-2d}\prod_{(i,j)\in\kappa}h_{ij}^2,$$

where h_{ij} is the hook-length of κ at (i,j):

$$h_{ij}=\kappa_i+\kappa_j'-i-j+1.$$

Let f_κ denote the number of standard tableau of shape κ [Ma1, p. 5]. The hook formula gives

$$f_\kappa = \frac{d!}{\prod_{(i,j)\in\kappa} h_{ij}}.$$

The asymptotic behavior of f_κ has been determined by Regev [R], which implies

$$f_\kappa \le C_1 m^d,$$

where C_1 is a constant depending only on m. Hence

$$C_1^{-1} m^{-d} d! \le \prod_{(i,j)\in\kappa} h_{ij},$$

so that we have

$$j_\kappa \ge C_1^{-2}(\beta m)^{-2d}(d!)^2. \tag{23}$$

Let h_d denote the dth *complete symmetric function*: $h_d = \sum_{|\nu|=d} m_\nu$. Then we have $h_d = \sum_{|\nu|=d} z_\nu^{-1} p_\nu$ (see [Ma1, p. 171]). Hence

$$\left|\sum_{|\nu|=d} \frac{p_\nu^2}{z_\nu \alpha^{l(\nu)}}\right| \le \|t\|^{2d}\binom{m+d-1}{d}(\beta m)^d. \tag{24}$$

Stirling's formula gives $\binom{m+d-1}{d} \le C_2 \cdot d^m$, C_2 a constant depending only on m. Therefore, (21) is a consequence of (22), (23), and (24).

Proposition 1. (1) *If $p \le q$, then the series* (20) *converges absolutely for all $t \in \mathbb{C}^m$.*

(2) *If $p = q+1$, then* (20) *converges absolutely for $\|t\| < \rho$ for some positive constant ρ.*

(3) *If $p > q+1$, then* (20) *diverges unless it terminates.*

Proof. We compare the series (20) with the generalized hypergeometric series

$${}_pF_q(a_1, \ldots, a_p;\, b_1, \ldots, b_q;\, z) = \sum_{d=0}^{\infty} \frac{(a_1)_d \cdots (a_p)_d}{(b_1)_d \cdots (b_q)_d} \frac{z^d}{d!}, \tag{25}$$

which is known to have radius of convergence $\rho = \infty$ if $p \le q$, $\rho = 1$ if $p = q+1$, $\rho = 0$ if $p > q+1$ unless it terminates. Put

$$a_{ji} = a_j - \frac{1}{\alpha}(i-1), \qquad b_{ki} = b_k - \frac{1}{\alpha}(i-1), \quad 1 \le j \le p, \quad 1 \le k \le q, \quad 1 \le i \le m.$$

It follows from Lemma 1 that

$$|C_\kappa^{(\alpha)}(t)| \le C \cdot R^d \|t\|^d$$

for some R. Note also that

$$\kappa_1!\kappa_2! \cdots \kappa_m! \le d!, \qquad d = |\kappa|.$$

These inequalities imply

$$\begin{aligned}
&\sum_{d=0}^{\infty} \sum_{|\kappa|=d} \left| \frac{[a_1]_\kappa^{(\alpha)} \cdots [a_p]_\kappa^{(\alpha)}}{[b_1]_\kappa^{(\alpha)} \cdots [b_q]_\kappa^{(\alpha)}} \frac{C_\kappa^{(\alpha)}(t)}{d!} \right| \\
&\le C \cdot \sum_{d=0}^{\infty} \sum_{|\kappa|=d} \left| \frac{[a_1]_\kappa^{(\alpha)} \cdots [a_p]_\kappa^{(\alpha)}}{[b_1]_\kappa^{(\alpha)} \cdots [b_q]_\kappa^{(\alpha)}} \right| \frac{(R\|t\|)^d}{\kappa_1! \cdots \kappa_m!} \\
&\le C \cdot \prod_{i=1}^{m} \left\{ \sum_{\kappa_i=0}^{\infty} \left| \frac{(a_{1i})_{\kappa_i} \cdots (a_{pi})_{\kappa_i}}{(b_{1i})_{\kappa_i} \cdots (b_{qi})_{\kappa_i}} \right| \frac{(R\|t\|)^{\kappa_i}}{\kappa_i!} \right\}.
\end{aligned}$$

Thus parts (1) and (2) of the theorem are clear (put $\rho = R^{-1}$ in Case (2)).

For the proof of divergence in the case of (3), note first that if $\kappa=(d)=(d,0,\dots,0)$ and $t=1^r$, then (15) and (18) give

$$C_d^{(\alpha)}(1^r)=\left\{\prod_{i=1}^{d}(1+\alpha(i-1))\right\}^{-1}\left\{\prod_{i=1}^{d}(r+\alpha(i-1))\right\},$$

where $C_d^{(\alpha)}(t)$ is short for $C_{(d)}^{(\alpha)}(t)$. Suppose $t\neq(0,\dots,0)$ and put $\tau=\min_{1\leq i\leq m,t_i\neq 0}\{|t_i|\}$, $r=$ the number of nonzero t_i. Then we have

$$\begin{aligned}\sum_{d=0}^{\infty}\sum_{|\kappa|=d}&\left|\frac{[a_1]_\kappa^{(\alpha)}\cdots[a_p]_\kappa^{(\alpha)}}{[b_1]_\kappa^{(\alpha)}\cdots[b_q]_\kappa^{(\alpha)}}\frac{C_\kappa^{(\alpha)}(t)}{d!}\right|\\ &\geq\sum_{d=0}^{\infty}\left|\frac{(a_1)_d\cdots(a_p)_d}{(b_1)_d\cdots(b_q)_d}\right|\frac{\tau^d}{d!}\cdot C_d^{(\alpha)}(1^r).\end{aligned}$$

Since $C_d^{(\alpha)}(1^r)\geq 1$, we get the divergence.

4. Holonomic system for the hypergeometric function.

4.1. Uniqueness theorem. The following uniqueness property was first proved in the case $\alpha=2$ by Muirhead [Mu]. His proof can be easily generalized for general α.

THEOREM 2. *Assume that $c-(1/\alpha)(i-1)$ is not a negative integer or zero for $1\leq i\leq m$. Then each of the m differential equations in the system* (9) *has the same unique formal power series solution $F(t)$ subject to the following conditions*:

(a) *$F(t)$ is a symmetric function of $t_1,\dots,t_m$; and*

(b) *$F(t)$ has a formal power series expansion at $(0,\dots,0)$ with $F(0)=1$.*

Proof. Since we assume that $F(t)$ is symmetric, it is sufficient to consider the first differential equation of (9) $(i=1)$

$$\begin{aligned}&t_1(1-t_1)\frac{\partial^2F}{\partial t_1^2}+\left\{c-\frac{1}{\alpha}(m-1)-\left(a+b+1-\frac{1}{\alpha}(m-1)\right)t_1\right\}\frac{\partial F}{\partial t_1}-abF\\ &\quad+\frac{1}{\alpha}\left\{\sum_{j=2}^{m}\frac{t_1(1-t_1)}{t_1-t_j}\frac{\partial F}{\partial t_1}-\sum_{j=2}^{m}\frac{t_j(1-t_j)}{t_1-t_j}\frac{\partial F}{\partial t_j}\right\}=0.\end{aligned}\tag{26}$$

As in James [J] we transform (26) to a partial differential equation in terms of the elementary symmetric function $r_1,\dots,r_m$ of $t_1,\dots,t_m$. Let $r_j^{(i)}$ for $j=1,\dots,m-1$ denote the jth elementary symmetric function formed from the variables $t_1,\dots,t_m$ omitting t_i. Then clearly

$$r_j=t_ir_{j-1}^{(i)}+r_j^{(i)},\qquad j=2,\dots,m-1.\tag{27}$$

Introducing dummy variables

$$\begin{aligned}r_0&=r_0^{(i)}=0,\\ r_j&=0\qquad(j=-1,-2,\dots\text{ and }m+1,m+2,\dots),\\ r_j^{(i)}&=0\qquad(j=-1,-2,\dots\text{ and }m,m+1,\dots),\end{aligned}$$

we may extend the relations (27) to hold for $-\infty<j<\infty$. As

$$\frac{\partial}{\partial t_i}=\sum_{\nu=1}^{m}r_{\nu-1}^{(i)}\frac{\partial}{\partial r_\nu},$$

$$\frac{\partial^2}{\partial t_1^2}=\sum_{\mu,\nu=1}^{m}r_{\mu-1}^{(1)}r_{\nu-1}^{(1)}\frac{\partial^2}{\partial r_\mu\,\partial r_\nu},$$

the differential equation (26) becomes

$$\begin{aligned}&\sum_{\mu,\nu=1}^{m} t_1(1-t_1)r_{\mu-1}^{(1)}r_{\nu-1}^{(1)}\frac{\partial^2 F}{\partial r_\mu\,\partial r_\nu}\\ &\quad+\sum_{j=1}^{m}\left\{(c-(a+b+1)t_1)r_{j-1}^{(1)}+\frac{1}{\alpha}\left(\sum_{i=2}^{m}\frac{t_i(1-t_1)}{t_1-t_i}\,r_{j-1}^{(1)}-\sum_{i=2}^{m}\frac{t_i(1-t_i)}{t_1-t_i}\,r_{j-1}^{(i)}\right)\right\}\\ &\qquad\cdot\frac{\partial F}{\partial r_j}-abF=0.\end{aligned}\tag{28}$$

It follows from (27) that

$$\sum_{i=2}^{m}\left(\frac{t_i(1-t_1)}{t_1-t_i}\,r_{j-1}^{(1)}-\frac{t_i(1-t_i)}{t_1-t_i}\,r_{j=1}^{(i)}\right)=\sum_{i=2}^{m}\frac{t_i}{t_1-t_i}(r_{j-1}^{(1)}-r_{j-1}^{(i)}+r_j^{(1)}-r_j^{(i)})$$

and that

$$\sum_{i=2}^{m}(r_{j-1}^{(1)}-r_{j-1}^{(i)}+r_j^{(1)}-r_j^{(i)})=m(r_{j-1}^{(1)}+r_j^{(1)})-(m-j+1)r_{j-1}-(m-j)r_j.$$

Denoting by $r_j^{(1,\,i)}$ the jth elementary symmetric function in the variables $t_2,\dots,t_m$ omitting t_i $(i\neq 1)$, we have

$$\begin{aligned}&\sum_{i=2}^{m}\frac{t_1}{t_1-t_i}(r_{j-1}^{(1)}-r_{j-1}^{(i)}+r_j^{(1)}-r_j^{(i)})\\ &\quad=\sum_{i=2}^{m}t_1(r_{j-2}^{(1,i)}+r_{j-1}^{(1,i)})\qquad (r_{-1}^{(1,i)}=0)\\ &\quad=-((m-j+1)(r_{j-1}-r_{j-1}^{(1)})+(m-j)(r_j-r_j^{(1)})).\end{aligned}$$

Hence

$$\begin{aligned}&(c-(a+b+1)t_1)r_{j-1}^{(1)}+\frac{1}{\alpha}\sum_{i=2}^{m}\left(\frac{t_i(1-t_1)}{t_1-t_i}\,r_{j-1}^{(1)}-\frac{t_i(1-t_i)}{t_1-t_i}\,r_{j-1}^{(i)}\right)\\ &\quad=c-\frac{1}{\alpha}(j-1)r_{j-1}^{(1)}+\left(a+b+1-\frac{1}{\alpha}j\right)r_j^{(1)}-(a+b+1)r_j.\end{aligned}$$

On the other hand, we see that

$$\begin{aligned}t_1r_{\mu-1}^{(1)}r_{\nu-1}^{(1)}&=r_\mu r_{\nu-1}^{(1)}-r_\mu^{(1)}r_{\nu-1}^{(1)}\\ &=r_\mu r_{\nu-1}^{(1)}-r_\mu^{(1)}(r_{\nu-1}-t_1r_{\nu-2}^{(1)})\\ &=r_\mu r_{\nu-1}^{(1)}-r_\mu^{(1)}r_{\nu-1}+t_1r_\mu^{(1)}r_{\nu-2}^{(1)}.\end{aligned}$$

Iterating this relation we have

$$\begin{aligned}t_1r_{\mu-1}^{(1)}r_{\nu-1}^{(1)}=r_\mu r_{\nu-1}^{(1)}-r_{\nu-1}r_\mu^{(1)}+r_{\mu+1}r_{\nu-2}^{(1)}-r_{\nu-2}r_{\mu+1}^{(1)}+\cdots\\ +r_{\mu+\nu-1}r_0^{(1)}-r_0r_{\mu+\nu-1}^{(1)}+t_1r_{\mu+\nu-1}^{(1)}r_{-1}^{(1)}\end{aligned}$$

and $r_{-1}^{(1)}=0$. Therefore the differential equation (28) becomes

$$\begin{aligned}&\sum_{\mu,\nu=1}^{m}\left\{\sum_{j=1}^{m}a_{\mu\nu}^{(j)}(r_{j-1}^{(1)}-r_j+r_j^{(1)})\right\}\frac{\partial^2 F}{\partial r_\mu\,\partial r_\nu}\\ &\quad+\sum_{j=1}^{m}\left\{\left(c-\frac{1}{\alpha}(j-1)\right)r_{j-1}^{(1)}+\left(a+b+1-\frac{1}{\alpha}j\right)r_j^{(1)}-(a+b+1)r_j\right\}\\ &\qquad\cdot\frac{\partial F}{\partial r_j}-abF=0,\end{aligned}\tag{29}$$

where $a_{\mu\nu}^{(j)} = a_{\nu\mu}^{(j)}$ and for $\mu \le \nu$,

$$a_{\mu\nu}^{(j)} = \begin{array}{ll} r_{\mu+\nu-j} & \text{for } 1 \le j \le \mu \\ 0 & \text{for } \mu < j \le \nu \\ -r_{\mu+\nu-j} & \text{for } \nu < j \le \mu+\nu \\ 0 & \text{for } \mu+\nu < j \qquad j = 1, \dots, m. \end{array}$$

In (29) we can equate coefficients of $r_{j-1}^{(1)}$ to zero for $j = 1, \dots, m$ according to [J, Lemma, p. 371] to obtain the system of partial differential equations $(a_{\mu\nu}^{(0)} = 0)$:

$$\begin{aligned} &\sum_{\mu,\nu=1}^{m} (a_{\mu\nu}^{(j-1)} + a_{\mu\nu}^{(j)}) \frac{\partial^2 F}{\partial r_\mu \, \partial r_\nu} + \left(c - \frac{1}{\alpha}(j-1)\right) \frac{\partial F}{\partial r_j} + \left(a+b+1-\frac{1}{\alpha}(j-1)\right) \frac{\partial F}{\partial r_{j-1}} \\ &- \delta_{1j} \left\{ \sum_{\mu,\nu=1}^{m} \left(\sum_{i=1}^{m} a_{\mu\nu}^{(i)} r_i \right) \frac{\partial^2 F}{\partial r_\mu \, \partial r_\nu} + (a+b+1) \sum_{i=1}^{m} r_i \frac{\partial F}{\partial r_i} + abF \right\} = 0, \qquad j = 1, \dots, m. \end{aligned} \tag{30}$$

Now we put

$$F(r_1, \dots, r_m) = \sum_{j_1, \dots, j_m = 0}^{\infty} c(j_1, \dots, j_m) r_1^{j_1} \cdots r_m^{j_m} \tag{31}$$

with $c(0, \dots, 0) = 1$. Order the coefficients $c(j_1, \dots, j_m)$ in lexicographic ordering, counting j_m as the first letter and j_1 as the last. Substituting (31) in the differential equation (30) with $j = k$ and putting $r_{k+1} = \cdots = r_m = 0$, we obtain a recurrence relation which expresses $j_k(j_k - 1 + c - (1/\alpha)(k-1))c(j_1, \dots, j_k, 0, \dots, 0)$ in terms of coefficients of lower order. Since $j_k - 1 + c - (1/\alpha)(k-1)$ is not zero by assumption, one can iterate this reduction until one reaches $c(0, \dots, 0)$ which we put equal to one. Hence all the coefficients $c(j_1, \dots, j_m)$ of (31) are uniquely determined by the recurrence relations. This completes the proof of Theorem 2.

Remark. As a matter of fact the series (31) is absolutely convergent in a neighborhood of the origin. We will show that the series (31) is nothing but the hypergeometric series ${}_2F_1^{(\alpha)}(a, b; c; t)$ (Theorem 4 in § 4.3).

As was noted by Muirhead [Mu, p. 995], the coefficients $c(j_1, \dots, j_m)$ are functions of a, b, c, α, and j_i and are independent of the dimension m in the sense that $c(j_1, \dots, j_k, 0, \dots, 0) = c(j_1, \dots, j_k)$. In fact the coefficients in the system (30) do not involve m explicitly, so that the recurrence relations determining $c(j_1, \dots, j_k, 0, \dots, 0)$ and those of $c(j_1, \dots, j_k)$ are the same. We rearrange the series (31) as a series of Jack polynomials:

$$F(t) = \sum_{d=0}^{\infty} \sum_{|\kappa|=d} a_\kappa C_\kappa^{(\alpha)}(t). \tag{32}$$

The expressions of power products of the elementary symmetric functions by monomial symmetric functions m_κ and those of m_κ by the Jack polynomials do not explicitly depend on m by virtue of [Mal, (2.3), p. 13] and the definition of Jack symmetric functions. Hence a_κ are independent of m, i.e., $a_{(\kappa_1, \dots, \kappa_h, 0, \dots, 0)} = a_{(\kappa_1, \dots, \kappa_h)}$.

COROLLARY 1. *The solution $F(t)$ in Theorem 2 can be obtained as a series of Jack polynomials $F(t) = \sum_{d=0}^{\infty} \sum_{|\kappa|=d} a_\kappa C_\kappa^{(\alpha)}(t)$ with coefficients a_κ independent of m.*

4.2. Properties of Jack polynomials. For later use, we prepare some properties of Jack polynomials following [Mu] closely in the case $\alpha = 2$.

The *generalized binomial coefficient* $\binom{\kappa}{\sigma}$ is defined as the coefficient of $C_\sigma^{(\alpha)}(t)/C_\sigma^{(\alpha)}(1^m)$ in the binomial expansion

$$\frac{C_\kappa^{(\alpha)}(1+t_1,\dots,1+t_m)}{C_\kappa^{(\alpha)}(1^m)}=\sum_{s=0}^{|\kappa|}\sum_{|\sigma|=s}\binom{\kappa}{\sigma}\frac{C_\sigma^{(\alpha)}(t)}{C_\sigma^{(\alpha)}(1^m)}. \tag{33}$$

Denote $\sigma\subseteq\kappa$ if $\sigma_i\le\kappa_i$ for any i. We first state crucial properties of generalized binomial coefficients whose proofs were not given in [Mu].

THEOREM 3. (a) $\binom{\kappa}{\sigma}=0$ *unless* $\sigma\subseteq\kappa$.

(b) *The generalized binomial coefficients* $\binom{\kappa}{\sigma}$ *are independent of the dimension* m.

We leave the proof to § 5. As in [Mu] we introduce the following differential operators:

$$E=\sum_{i=1}^m t_i\frac{\partial}{\partial t_i},\qquad \varepsilon=\sum_{i=1}^m\frac{\partial}{\partial t_i},$$

$$\delta(\alpha)=\frac{\alpha}{2}\sum_{i=1}^m t_i\frac{\partial^2}{\partial t_i^2}+\sum_{\substack{1\le i,j\le m\\ i\ne j}}\left[\frac{t_i}{t_i-t_j}\right]\frac{\partial}{\partial t_i}.$$

First, by Euler's identity,

$$EC_\kappa^{(\alpha)}(t)=|\kappa|C_\kappa^{(\alpha)}(t). \tag{34}$$

For the partition κ, we put $\kappa_{(i)}=(\kappa_1,\kappa_2,\dots,\kappa_i+1,\dots,\kappa_m)$ and $\kappa^{(i)}=(\kappa_1,\kappa_2,\dots,\kappa_i-1,\dots,\kappa_m)$ and call them admissible if the parts are in nonincreasing order. Observe that if $|\kappa|=|\sigma|+1$, then $\binom{\kappa}{\sigma}=0$ unless $\sigma=\kappa^{(i)}$ for some i. We have

$$\frac{\varepsilon C_\kappa^{(\alpha)}(t)}{C_\kappa^{(\alpha)}(1^m)}=\sum_{i=1}^m\binom{\kappa}{\kappa^{(i)}}\frac{C_{\kappa^{(i)}}^{(\alpha)}(t)}{C_{\kappa^{(i)}}^{(\alpha)}(1^m)}, \tag{35}$$

$$\frac{\delta(\alpha)C_\kappa^{(\alpha)}(t)}{C_\kappa^{(\alpha)}(1^m)}=\sum_{i=1}^m\binom{\kappa}{\kappa^{(i)}}\left(\frac{\alpha}{2}(\kappa_i-1)+\frac{1}{2}(m-i)\right)\frac{C_{\kappa^{(i)}}^{(\alpha)}(t)}{C_{\kappa^{(i)}}^{(\alpha)}(1^m)}. \tag{36}$$

The summations in (35) and (36) are over all i such that $\kappa^{(i)}$ is admissible. This convention will be used in all future summations involving $\kappa_{(i)}$ and $\kappa^{(i)}$. Equation (35) can be proved in the following way:

$$\begin{aligned}\frac{\varepsilon C_\kappa^{(\alpha)}(t)}{C_\kappa^{(\alpha)}(1^m)}&=\sum_{i=1}^m\frac{\partial C_\kappa^{(\alpha)}(t)}{\partial t_i}C_\kappa^{(\alpha)}(1^m)\\ &=\lim_{\lambda\to0}\frac{C_\kappa^{(\alpha)}(t_1+\lambda,\dots,t_m+\lambda)-C_\kappa^{(\alpha)}(t)}{\lambda C_\kappa^{(\alpha)}(1^m)}\\ &=\lim_{\lambda\to0}\sum_{i=1}^m\binom{\kappa}{\kappa^{(i)}}\frac{C_{\kappa^{(i)}}^{(\alpha)}(t)}{C_{\kappa^{(i)}}^{(\alpha)}(1^m)}+\text{terms of higher degree in }\lambda\\ &=\sum_{i=1}^m\binom{\kappa}{\kappa^{(i)}}\frac{C_{\kappa^{(i)}}^{(\alpha)}(t)}{C_{\kappa^{(i)}}^{(\alpha)}(1^m)}.\end{aligned}$$

Equation (36) follows by noting that $\delta(\alpha)=\frac{1}{2}(\varepsilon D(\alpha)-D(\alpha)\varepsilon)$ and by applying the operators ε and $D(\alpha)$ to $C_\kappa^{(\alpha)}(t)/C_\kappa^{(\alpha)}(1^m)$.

Next we show that

$$p_1\exp(t_1+\cdots+t_m)=\frac{\sum_{d=0}^\infty\sum_{|\kappa|=d}dC_\kappa^{(\alpha)}}{d!}, \tag{37}$$

$$\text{(38)} \qquad \frac{\alpha}{2} p_2 \exp(t_1+\cdots+t_m) = \sum_{d=0}^{\infty} \sum_{|\kappa|=d} \left(\rho_\kappa(\alpha)+\left(1-\frac{\alpha}{2}\right)d\right) \frac{C_\kappa^{(\alpha)}(t)}{d!},$$

$$\text{(39)} \qquad \frac{\alpha}{2} p_1 p_2 \exp(t_1+\cdots+t_m) = \sum_{d=0}^{\infty} \sum_{|\kappa|=d} (d-2)\left(\rho_\kappa(\alpha)+\left(1-\frac{\alpha}{2}\right)d\right) \frac{C_\kappa^{(\alpha)}(t)}{d!},$$

where $p_i = t_1^i+\cdots+t_m^i$ and $\rho_\kappa(\alpha)=\sum_{i=1}^m [\kappa_i((\alpha/2)\kappa_i - i)]$. Note that $e_\kappa(\alpha)=\rho_\kappa(\alpha)+(m-(\alpha/2))|\kappa|$. Equation (37) follows from (17) since

$$\begin{aligned} p_1 \exp(t_1+\cdots+t_m) &= \sum_{d=0}^{\infty} \frac{p_1^{d+1}}{d!} = \sum_{d=0}^{\infty} \frac{dp_1^d}{d!} \\ &= \sum_{d=0}^{\infty} \sum_{|\kappa|=d} \frac{dC_\kappa^{(\alpha)}(t)}{d!}. \end{aligned}$$

Applying $D(\alpha)$ to both sides of

$$\text{(40)} \qquad \exp(t_1+\cdots+t_m) = \sum_{d=0}^{\infty} \sum_{|\kappa|=d} \frac{C_\kappa^{(\alpha)}(t)}{d!}$$

and equating coefficients of $C_\kappa^{(\alpha)}(t)$ using (12) and (13) gives (38). Applying E to both sides of (38) and equating coefficients of $C_\kappa^{(\alpha)}$ using (38) gives (39). Using these formulas we obtain

$$\text{(41)} \qquad \sum_{i=1}^{m} \binom{\kappa_{(i)}}{\kappa} C_{\kappa_{(i)}}^{(\alpha)}(1^m) = m(|\kappa|+1)C_\kappa^{(\alpha)}(1^m),$$

$$\text{(42)} \qquad \sum_{i=1}^{m} \binom{\kappa_{(i)}}{\kappa}\left(\kappa_i - \frac{1}{\alpha}(i-1)\right) C_{\kappa_{(i)}}^{(\alpha)}(1^m) = |\kappa|(|\kappa|+1)C_\kappa^{(\alpha)}(1^m),$$

$$\text{(43)} \qquad \begin{aligned} &\sum_{i=1}^{m} \binom{\kappa_{(i)}}{\kappa}\left(\kappa_i - \frac{1}{\alpha}(i-1)\right)^2 C_{\kappa_{(i)}}^{(\alpha)}(1^m) \\ &\quad = \frac{2}{\alpha}(|\kappa|+1)\left(\rho_\kappa(\alpha)+\frac{1}{2}(m+1)|\kappa|\right)C_\kappa^{(\alpha)}(1^m). \end{aligned}$$

Applying ε to both sides of (40) and equating coefficients of $C_\kappa^{(\alpha)}(t)$ gives (41). Applying $\delta(\alpha)$ to both sides of (37) and collecting coefficients of $C_\kappa^{(\alpha)}(t)$ using (37) gives (42). Applying $\delta(\alpha)$ to both sides of (38) and collecting coefficients of $C_\kappa^{(\alpha)}(t)$ using (37), (38), (39), (41), and (42) gives (43).

4.3. Holonomic system. We want to show that the hypergeometric function ${}_2F_1^{(\alpha)}(a, b; c; t)$ satisfies the holonomic system (9). For this, as in [Mu], we will deal with the differential equation formed by summing the differential equations in the system (9):

$$\text{(44)} \qquad \begin{aligned} &\sum_{i=1}^{m} t_i(1-t_i)\frac{\partial^2 F}{\partial t_i^2} + \sum_{i=1}^{m}\left\{c - \frac{1}{\alpha}(m-1) - \left(a+b+1-\frac{1}{\alpha}(m-1)\right)t_i\right\} \\ &\quad \cdot \frac{\partial F}{\partial t_i} - mabF + \frac{2}{\alpha}\sum_{\substack{1\le i,j\le m \\ i\ne j}} \frac{t_i(1-t_i)}{t_i - t_j}\frac{\partial F}{\partial t_i} = 0. \end{aligned}$$

LEMMA 2. *Assume that $c-(1/\alpha)(i-1)$ is not a negative integer or zero for $1 \le i \le m$. Then the solution $F(t)$ of (44) of the form*

$$\text{(45)} \qquad F(t) = \sum_{d=0}^{\infty} \sum_{|\kappa|=d} \gamma_\kappa C_\kappa^{(\alpha)}(t), \qquad \gamma_{(0)} = 1$$

is unique if the coefficients γ_κ are independent of m.

Proof. Substituting (45) in (44) and using (12), (34), (35), and (36), we obtain the following recurrence relations for γ_κ:

$$\begin{aligned}(46)\qquad &\sum_{i=1}^{m}\binom{\kappa_{(i)}}{\kappa}\left\{\frac{\alpha}{2}(c+\kappa_i)-\frac{1}{2}(i-1)\right\}C^{(\alpha)}_{\kappa_{(i)}}(1^m)\gamma_{\kappa_{(i)}}\\ &=\left\{\frac{\alpha}{2}mab+\rho_\kappa(\alpha)+\frac{\alpha}{2}(a+b)|\kappa|+\frac{1}{2}(m+1)|\kappa|\right\}C^{(\alpha)}_{\kappa}(1^m)\gamma_\kappa.\end{aligned}$$

Put $\beta_\kappa=\alpha^{|\kappa|}|\kappa|!\,j_\kappa^{-1}[c]^{(\alpha)}_\kappa\gamma_\kappa$. Then, in virtue of (18), (46) becomes

$$\begin{aligned}(47)\qquad &\sum_{i=1}^{m}\binom{\kappa_{(i)}}{\kappa}(m+\alpha\kappa_i-i+1)\beta_{\kappa_{(i)}}\\ &=\left\{\frac{\alpha}{2}mab+\rho_\kappa(\alpha)+\frac{\alpha}{2}(a+b)|\kappa|+\frac{1}{2}(m+1)|\kappa|\right\}\beta_\kappa.\end{aligned}$$

Now assume that γ_κ and hence β_κ are independent of m. For an arbitrary partition κ we take $m>l(\kappa)$ so that $\kappa_{(m+1)}$ is not admissible. Then, on replacing m in (47) with $m+1$, we have

$$\begin{aligned}(48)\qquad &\sum_{i=1}^{m}\binom{\kappa_{(i)}}{\kappa}(m+\alpha\kappa_i-i+2)\beta_{\kappa_{(i)}}\\ &=\left\{\frac{\alpha}{2}(m+1)ab+\rho_\kappa(\alpha)+\frac{\alpha}{2}(a+b)|\kappa|+\frac{1}{2}(m+2)|\kappa|\right\}\beta_\kappa.\end{aligned}$$

Subtracting each side of (46) from that of (48) yields

$$(49)\qquad \sum_{i}\binom{\kappa_{(i)}}{\kappa}\beta_{\kappa_{(i)}}=\left(\frac{\alpha}{2}ab+\frac{|\kappa|}{2}\right)\beta_\kappa,$$

and equating constant terms of (47) gives

$$(50)\qquad \sum_{i}\binom{\kappa_{(i)}}{\kappa}(\alpha\kappa_i-i+1)\beta_{\kappa_{(i)}}=\left\{\rho_\kappa(\alpha)+\frac{\alpha}{2}|\kappa|(a+b)+\frac{|\kappa|}{2}\right\}\beta_\kappa.$$

As κ runs over all partitions of d, (49) and (50) give equations in all the unknown β corresponding to partitions of $d+1$, since any partition of $d+1$ can be expressed as $\kappa_{(i)}$ for some i and some partition κ of d. For partitions κ and ν with $|\nu|=|\kappa|$, write $\nu\leq^R\kappa$ (reverse lexicographic order) if $\nu=\kappa$ or the first nonvanishing difference $\kappa_i-\nu_i$ is positive. We prove that (49) and (50) determine β_κ uniquely ($\beta_{(0)}=1$) by induction on $\leq^R$ of indices κ. Suppose that the β_ν with $|\nu|<|\kappa|$ or $\nu\nleq^R\kappa$ have been determined and that $\kappa_1=\cdots=\kappa_r>\kappa_{r+1}$. Put $\sigma=\kappa^{(r)}$ and replace κ in (49) and (50) by σ. Then, as

$$\binom{\sigma_{(r)}}{\sigma}=\binom{\kappa}{\sigma}\neq 0,$$

it is readily verified that (49) and (50) afford an explicit formula of β_κ in terms of lower-order ones ((49) alone will do in case $r=1$). Thus β_κ are uniquely determined.

LEMMA 3. *${}_2F_1^{(\alpha)}(a,b;c;t)$ is a solution of the differential equation* (44).

Proof. It suffices to show that

$$(51)\qquad \gamma_\kappa=\frac{[a]^{(\alpha)}_\kappa[b]^{(\alpha)}_\kappa}{[c]^{(\alpha)}_\kappa|\kappa|!}$$

satisfies the recurrence relations (46). Substituting (51) into (46), the problem reduces to showing that

$$\sum_{i=1}^{m}\binom{\kappa_{(i)}}{\kappa}\left(a+\kappa_i-\frac{1}{\alpha}(i-1)\right)\left(b+\kappa_i-\frac{1}{\alpha}(i-1)\right)C^{(\alpha)}_{\kappa_{(i)}}(1^m)$$
$$=\frac{2}{\alpha}(|\kappa|+1)\left(\frac{\alpha}{2}mab+\rho_\kappa(\alpha)+\frac{\alpha}{2}(a+b)|\kappa|+\frac{1}{2}(m+1)|\kappa|\right)C^{(\alpha)}_{\kappa}(1^m).$$

But this follows at once from (41), (42), and (43).

As the coefficients γ_κ given by (51) are clearly independent of m, it follows from Lemmas 2 and 3 that the hypergeometric function ${}_2F_1^{(\alpha)}(a, b; c; t)$ is the unique solution of (44) among the formal series of form (45) with γ_κ being independent of m. On the other hand, the unique formal series solution $F(t)$ of (9) given by Theorem 2 also satisfies (44) and is of the form (45) with γ_κ being independent of m. Therefore, we get the desired equality $F(t)={}_2F_1^{(\alpha)}(a, b; c; t)$.

We summarize our results in the following theorem.

THEOREM 4. *${}_2F_1^{(\alpha)}(a, b; c; t_1, \ldots, t_m)$ is the unique solution of each of the m differential equations in the system* (9) *subject to the following conditions*:

(a) *$F(t)$ is a symmetric function of $t_1, \ldots, t_m$*; and

(b) *$F(t)$ is analytic at the origin with $F(0)=1$.*

5. Proof of Theorem 3.

5.1. Preliminaries. Define rational functions $g^\lambda_{\mu\nu}=g^\lambda_{\mu\nu}(\alpha)$ by

$$J_\mu(t)J_\nu(t)=\sum_\lambda j_\lambda^{-1}g^\lambda_{\mu\nu}(\alpha)J_\lambda(t).$$

Note that $g^\lambda_{\mu\nu}=0$ unless $|\lambda|=|\mu|+|\nu|$, $\mu\subseteq\lambda$ and $\nu\subseteq\lambda$ [St, Corollary 6.4]. If $\mu\subseteq\lambda$ then the skew shape λ/μ (regarded as a difference $\lambda-\mu$ of diagrams) is called a *horizontal strip* if no two different points of λ/μ lie in the same column. Call a horizontal strip λ/μ an *n-strip* if $|\lambda/\mu|=n$. Then $g^\lambda_{\mu n}=g^\lambda_{\mu(n)}\neq 0$ if and only if $\mu\subseteq\lambda$ and λ/μ is a horizontal n-strip [St, Prop. 5.3]. Moreover, $g^\lambda_{\mu n}$ has the explicit evaluation [St, Thm. 6.1]:

$$g^\lambda_{\mu n}=n!\,\alpha^n\left(\prod_{s\in\mu}A_{\lambda\mu}(s)\right)\left(\prod_{s\in\lambda}B_{\lambda\mu}(s)\right), \tag{52}$$

where

$$A_{\lambda\mu}(s)=\begin{cases}h^\mu_*(s), & \text{if } \lambda/\mu \text{ does not contain a square in the same column as } s,\\ h^*_\mu(s), & \text{otherwise,}\end{cases}$$

$$B_{\lambda\mu}(s)=\begin{cases}h^*_\lambda(s), & \text{if } \lambda/\mu \text{ does not contain a square in the same column as } s,\\ h^\lambda_*(s), & \text{otherwise.}\end{cases}$$

We will use only the case $n=1$ of this formula.

Let $x=(x_1, x_2, \ldots)$ be other variables. Then we have [St, Prop. 4.2]:

$$J_\lambda(t, x)=\sum_{\mu,\nu}g^\lambda_{\mu\nu}(\alpha)J_\mu(t)j_\mu^{-1}J_\nu(x)j_\nu^{-1}. \tag{53}$$

We will also need some explicit formulas. Let $l(\lambda)=m$ and write $\lambda-I=\lambda-I_m=(\lambda_1-1, \lambda_2-1, \ldots, \lambda_m-1)$. Then there holds [St, Props. 5.1 and 5.5]

$$J_\lambda(t_1, \ldots, t_m)=\prod_{i=1}^{m}h^\lambda_*(i, 1)t_1\cdots t_mJ_{\lambda-I}(t_1, \ldots, t_m). \tag{54}$$

Consequently, $v_{\lambda\lambda}(\alpha)$ is given by [St, Thm. 5.6]:

$$v_{\lambda\lambda}(\alpha)=\prod_{s\in\lambda} h_*^\lambda(s). \tag{55}$$

In this section we will denote the generalized binomial coefficient by $\binom{\kappa}{\sigma}_m$ to express its apparent dimensional dependence in the definition (33).

We proceed via a sequence of lemmas.

Lemma 4. *For fixed d with $|\sigma|\le d\le|\kappa|$, we have*

$$\binom{|\kappa|-|\sigma|}{d-|\sigma|}\binom{\kappa}{\sigma}_m=\sum_{|\tau|=d}\binom{\kappa}{\tau}_m\binom{\tau}{\sigma}_m .$$

Proof. Let s be an indeterminate. By definition of $\binom{\kappa}{\sigma}_m$ we have

$$\begin{aligned}\frac{J_\kappa(1+s+t_1,\dots,1+s+t_m)}{J_\kappa(1^m)}&=\sum_\tau\binom{\kappa}{\tau}_m\frac{J_\tau(s+t_1,\dots,s+t_m)}{J_\tau(1^m)}\\&=\sum_\tau\binom{\kappa}{\tau}_m\left\{\sum_\sigma s^{|\lambda|-|\sigma|}\binom{\tau}{\sigma}_m\frac{J_\sigma(t)}{J_\sigma(1^m)}\right\},\end{aligned}$$

while we see that

$$\frac{J_\kappa(1+s+t_1,\dots,1+s+t_m)}{J_\kappa(1^m)}=\frac{\sum_\sigma(1+s)^{|\kappa|-|\sigma|}\binom{\kappa}{\sigma}_m J_\sigma(t)}{J_\sigma(1^m)}.$$

Hence equating the coefficients of $s^{d-|\sigma|}J_\sigma(t)$ of each of the right sides of these equalities gives the desired result.

Lemma 5. *Let μ and ν be partitions with $|\mu|=|\nu|$ and of lengths $\le m$. Then*

$$\begin{aligned}\sum_\tau J_\tau(1^m)^{-1}g_{\tau 1}^{\nu}\binom{\mu}{\tau}_m&=J_\mu(1^m)^{-1}J_\nu(1^m)^{-1}j_\nu\\&\quad\cdot\sum_\rho J_\rho(1^m)j_\rho^{-1}g_{\mu 1}^{\rho}\binom{\rho}{\nu}_m-mJ_\nu(1^m)^{-1}\binom{\mu}{\nu}_m .\end{aligned}\tag{56}$$

Proof. Multiplying both sides of (33) (in which we replace κ and λ with μ and ν, respectively) by p_1 yields

$$\frac{p_1J_\mu(1+t_1,\dots,1+t_m)}{J_\mu(1^m)}=\sum_\tau\binom{\mu}{\tau}_m J_\tau(1^m)^{-1}\left\{\sum_\nu j_\nu^{-1}g_{\tau 1}^{\nu}J_\nu(t)\right\}.\tag{57}$$

As $p_1=1+t_1+\cdots+1+t_m-m$ we see that

$$\begin{aligned}&\frac{p_1J_\mu(1+t_1,\dots,1+t_m)}{J_\mu(1^m)}\\&=J_\mu(1^m)^{-1}\left\{\sum_\rho j_\rho^{-1}g_{\mu 1}^{\rho}J_\rho(1+t_1,\dots,1+t_m)-mJ_\mu(1+t_1,\dots,1+t_m)\right\}\\&=J_\mu(1^m)^{-1}\sum_\rho j_\rho^{-1}g_{\mu 1}^{\rho}J_\rho(1^m)\left\{\sum_\nu\binom{\rho}{\nu}_m\frac{J_\nu(t)}{J_\nu(1^m)}\right\}-m\sum_\tau\binom{\mu}{\sigma}_m\frac{J_\tau(t)}{J_\tau(1^m)}.\end{aligned}\tag{58}$$

Then equating the coefficients of $J_\nu(t)$ in (57) and (58) gives (56).

Note that if $|\mu|=|\nu|$ and $\mu\ne\nu$, then $\binom{\mu}{\nu}_m=0$ and that $\binom{\mu}{\mu}_m=1$.

By Lemma 4, for the proof of Theorem 3, it will suffice to prove the case $|\kappa|=|\sigma|+1$ of the theorem, which we assume henceforth. For a partition $\sigma=(\sigma_1,\sigma_2,\dots)$, write $\sigma_-=(\sigma_1,\sigma_2,\dots,\sigma_{m-1})$.

LEMMA 6. *Suppose that* (a) *holds in the dimensions* $\leq m-1$ *and that* $\sigma_- \not\subseteq \kappa$. *Then* $\binom{\kappa}{\sigma}_m = 0$.

Proof. Since $j_d^{-1}J_d(t_m) = (\alpha^d d!)^{-1}t_m^d$ by (15) and (18), (53) implies that

$$\begin{aligned} J_\kappa(1+t_1, \ldots, 1+t_m) &= \sum_{\mu,d} (\alpha^d d!)^{-1} g_{\mu d}^\kappa j_\mu^{-1} J_\mu(1+t_1, \ldots, 1+t_{m-1})(1+t_m)^d \\ &= \sum_{\substack{\mu,\lambda,d \\ \lambda\subseteq\mu}} (\alpha^d d!)^{-1} j_\mu^{-1} J_\mu(1^{m-1}) J_\lambda(1^{m-1})^{-1} \\ &\quad \cdot g_{\mu d}^\kappa \binom{\mu}{\lambda}_{m-1} J_\lambda(t_1, \ldots, t_{m-1})(1+t_m)^d, \\ J_\sigma(t_1, \ldots, t_m) &= \sum_{\nu,d} (\alpha^d d!)^{-1} g_{\nu d}^\sigma j_\nu^{-1} J_\nu(t_1, \ldots, t_{m-1}) t_m^d. \end{aligned}$$

Substituting these into (33), we get

$$\begin{aligned} &\sum_{\substack{\mu,\lambda,d \\ \lambda\subseteq\mu}} (\alpha^d d!)^{-1} j_\mu^{-1} J_\mu(1^{m-1}) J_\lambda(1^{m-1})^{-1} g_{\mu d}^\kappa \binom{\mu}{\lambda}_{m-1} J_\lambda(t_1, \ldots, t_{m-1})(1+t_m)^d \\ &= \sum_{\substack{\sigma,\nu,d \\ \nu\subseteq\sigma}} (\alpha^d d!)^{-1} j_\nu^{-1} J_\kappa(1^m) J_\sigma(1^m)^{-1} g_{\nu d}^\sigma \binom{\kappa}{\sigma}_m J_\nu(t_1, \ldots, t_{m-1}) t_m^d. \end{aligned} \tag{59}$$

Put $\nu = \sigma_-$ and $t_m = 1$ in (59). Then comparing the coefficients of $J_{\sigma_-}(t_1, \ldots, t_{m-1})$ to both sides gives

$$\sum_\tau J_\tau(1^m)^{-1} g_{\sigma_-\sigma_m}^\tau \binom{\kappa}{\tau}_m = 0.$$

Suppose first that $\sigma_m = 0$. Then τ must be identical with σ, so that $\binom{\kappa}{\sigma}_m = 0$. The general case follows using induction on σ_m.

LEMMA 7. *For a partition* $\lambda = l^p$, $p \leq m$, *we have*

$$\binom{\lambda}{\lambda^{(p)}}_m = pl.$$

Proof. As the proof of the case $p = m$ is clear, we assume $p < m$. By putting $t_{p+1} = \cdots = t_m = 0$ in (33) and using (53), we have

$$\sum_{\mu,\nu} g_{\mu\nu}^\lambda J_\mu(1+t_1, \ldots, 1+t_p) j_\mu^{-1} J_\nu(1^{m-p}) j_\nu^{-1} = J_\lambda(1^m) \sum_\sigma \binom{\lambda}{\sigma}_m \frac{J_\sigma(t_1, \ldots, t_p)}{J_\sigma(1^m)}.$$

Equating the coefficients of $m_{\lambda^{(p)}}(t_1, \ldots, t_p)$ of both sides yields

$$l g_{\lambda 0}^\lambda j_\lambda^{-1} v_{\lambda\lambda} + g_{\lambda^{(p)}1}^\lambda j_{\lambda^{(p)}}^{-1} v_{\lambda^{(p)},\lambda^{(p)}} J_1(1^{m-p}) j_1^{-1} = J_\lambda(1^m) J_{\lambda^{(p)}}^{-1}(1^m) v_{\lambda^{(p)},\lambda^{(p)}} \binom{\lambda}{\lambda^{(p)}}_m.$$

Thus the desired result follows at once if we substitute the explicit formulas of $g_{\nu n}^\lambda$, j_λ, $J_\lambda(1^r)$, and $v_{\lambda\lambda}$.

LEMMA 8. *We have*

$$\begin{aligned} v_{(r^p,1^s),(r^{p-1},r-1,1^{s+1})} &= p(r-1) \sum_{i=0}^s \left\{ \prod_{j=0}^{i-1} \frac{p+s-j+\alpha(r-1)}{s-j+\alpha(r-1)} (s-j) \right\} \frac{s-i+\alpha r}{s-i+\alpha(r-1)} \\ &\quad \cdot v_{(r^{p-1},r-1,1^{s-i}),(r^{p-1},r-1,1^{s-i})}. \end{aligned}$$

Proof. Equation (53) implies that

$$J_{(r^p,1^s)}(t_1, \ldots, t_{p+s+1}) = \sum_\lambda (\alpha^d d!)^{-1} g_{\lambda d}^{(r^p,1^s)} j_\lambda^{-1} J_\lambda(t_1, \ldots, t_{p+s}) t_{p+s+1}^d.$$

From this one can easily deduce that

$$\begin{aligned} v_{(r^p,1^s),(r^{p-1},r-1,1^{s+1})} = {} & \alpha^{-1} g^{(r^p,1^s)}_{(r^{p-1},r-1,1^s),1} \dot{j}^{-1}_{(r^{p-1},r-1,1^s)} v_{(r^{p-1},r-1,1^s),(r^{p-1},r-1,1^s)} \\ & + \alpha^{-1} g^{(r^p,1^s)}_{(r^p,1^{s-1}),1} \dot{j}^{-1}_{(r^p,1^{s-1})} v_{(r^p,1^{s-1}),(r^{p-1},r-1,1^s)}, \end{aligned}$$

where we see by (15) and (52) that

$$g^{(r^p,1^s)}_{(r^{p-1},r-1,1^s),1} \dot{j}^{-1}_{(r^{p-1},r-1,1^s)} = \alpha p(r-1) \frac{s+\alpha r}{s+\alpha(r-1)},$$

$$g^{(r^p,1^s)}_{(r^p,1^{s-1}),1} \dot{j}^{-1}_{(r^p,1^{s-1})} = \alpha s \frac{p+s+\alpha(r-1)}{s+\alpha(r-1)}.$$

Using this formula recursively, we arrive at the desired formula.

Proof of Theorem 3(a). We shall prove (a) of Theorem 3 by induction on the dimension m, the proof being clear for the case $m=1$. We assume that $l(\kappa)=m$. By Lemma 6 we may also assume $\sigma_- \subseteq \kappa$ and $\sigma \not\subseteq \kappa$. We proceed by induction on $d = \kappa_{m-1} - \kappa_m$, $d \geq 0$: Suppose that (a) holds in the cases $\kappa_{m-1} - \kappa_m \leq d-1$. The proof of the case $\kappa_{m-1} = \kappa_m$ is clear because $\sigma \not\subseteq \kappa$ implies $\sigma_- \not\subseteq \kappa$. Suppose $\sigma_m - \kappa_m \geq 2$ and put $\mu = \kappa^{(m-1)}$ and $\nu = \sigma$ in (56). Then one can easily verify by induction on d that all the terms of both sides of (56) except the one with $\binom{\kappa}{\sigma}_m$ vanish, so that $\binom{\kappa}{\sigma}_m = 0$. Hence it suffices to consider the case $\sigma_m = \kappa_m + 1$. Since $|\kappa| = |\sigma| + 1$, it follows that $\kappa_{m-1} - \sigma_{m-1} \leq 2$.

Case $\kappa_{m-1} = \sigma_{m-1}$. Put $\mu = \kappa^{(m-1)}$ and $\nu = \sigma$ in (56). Then by using Lemma 6 one can easily show that $\binom{\kappa}{\sigma}_m = 0$ if $d=1$. Similarly if $d \geq 2$, one can deduce that $\binom{\kappa}{\sigma}_m$ is proportional to

$$\binom{\kappa^{(m-1)}}{\sigma^{(m-1)}},$$

so that $\binom{\kappa}{\sigma}_m = 0$ by induction on d.

Case $\kappa_{m-1} = \sigma_{m-1} + 1$. It necessarily follows that $m \geq 3$ and $d \geq 2$. Note that $\kappa^{(m-1)}_{(m)} = \sigma_{(p)}$ for some $p \leq m-2$. Putting $\mu = \kappa^{(m-1)}$ and $\nu = \sigma$ in (56), we have

$$\begin{aligned} & J_{\sigma^{(m)}}(1^m)^{-1} g^{\sigma}_{\sigma^{(m)}1} \binom{\kappa^{(m-1)}}{\sigma^{(m)}}_m \\ & = \{J_{\kappa^{(m-1)}}(1^m) J_\sigma(1^m)\}^{-1} \dot{j}_\sigma \\ & \quad \cdot \left\{ J_{\kappa^{(m-1)}_{(m)}}(1^m) \dot{j}^{-1}_{\kappa^{(m-1)}_{(m)}} g^{\kappa^{(m-1)}_{(m)}}_{\kappa^{(m-1)}1} \binom{\kappa^{(m-1)}}{\sigma}_m + J_\kappa(1^m) \dot{j}_\kappa^{-1} g^{\kappa}_{\kappa^{(m-1)}1} \binom{\kappa}{\sigma}_m \right\}. \end{aligned}$$

Observe that (18) implies

$$J_\kappa(1^m) J_{\kappa^{(j)}}(1^m)^{-1} J_{\kappa^{(j)}_{(i)}}(1^m) J_{\kappa_{(i)}}(1^m)^{-1} = 1. \tag{60}$$

Hence the proof of $\binom{\kappa}{\sigma}_m = 0$ is equivalent to show that

$$\binom{\sigma_{(p)}}{\sigma}_m = \left(\dot{j}^{-1}_{\sigma_{(p)}} g^{\sigma_{(p)}}_{\sigma^{(m)}_{(p)}1} \right)^{-1} \dot{j}_\sigma^{-1} g^{\sigma}_{\sigma^{(m)}1} \binom{\sigma^{(m)}_{(p)}}{\sigma^{(m)}}_m. \tag{61}$$

Note first that by (15) and (52) we have

$$\begin{aligned} \dot{j}_\lambda^{-1} g^{\lambda}_{\lambda^{(p)}1} & = \prod_{j=1}^{\lambda_p - 1} h_*^{\lambda^{(p)}}(p,j) h_*^{\lambda}(p,j)^{-1} \cdot \prod_{i=1}^{p-1} h^*_{\lambda^{(p)}}(i, \lambda_p) h^*_\lambda(i, \lambda_p)^{-1} \\ & = \prod_{j=1}^{\lambda_p - 1} \frac{(\lambda'_j - p + 1 + \alpha(\lambda_p - j - 1))}{(\lambda'_j - p + 1 + \alpha(\lambda_p - j))} \cdot \prod_{i=1}^{p-1} \frac{(p - i - 1 + \alpha(\lambda_i - \lambda_p + 1))}{(p - i + \alpha(\lambda_i - \lambda_p + 1))}. \end{aligned} \tag{62}$$

Suppose $d\geq 3$, so that $\sigma_{m-1}>\sigma_m$. Then by induction on d we see that

$$\binom{\sigma_{(p)}^{(m-1)}}{\sigma^{(m-1)}}_m=\left(j^{-1}_{\sigma_{(p)}^{(m-1)}}g^{\sigma_{(p)}^{(m-1)}}_{\sigma_{(p)}^{(m-1,m)}1}\right)^{-1}j^{-1}_{\sigma^{(m-1)}}g^{\sigma^{(m-1)}}_{\sigma^{(m-1,m)}1}\binom{\sigma_{(p)}^{(m-1,m)}}{\sigma^{(m-1,m)}}_m, \tag{63}$$

where $\sigma^{(m-1,m)}=(\sigma_1,\sigma_2,\dots,\sigma_{m-1}-1,\sigma_m-1)$. Observe that the coefficients of

$$\binom{\sigma_{(p)}^{(m)}}{\sigma^{(m)}}_m \quad\text{and}\quad \binom{\sigma_{(p)}^{(m-1,m)}}{\sigma^{(m-1,m)}}_m$$

are the same in view of (62). Put $\mu=\sigma_{(p)}^{(m-1)}$ and $\nu=\sigma$ (respectively, $\mu=\sigma_{(p)}^{(m-1,m)}$, $\nu=\sigma^{(m)}$) in (56). Then by (60) we have

$$\binom{\sigma_{(p)}^{(m-1)}}{\sigma^{(m-1)}}_m=j_\sigma\left(g^{\sigma}_{\sigma^{(m-1)}1}\right)^{-1}j^{-1}_{\sigma_{(p)}}g^{\sigma_{(p)}}_{\sigma_{(p)}^{(m-1)}1}\binom{\sigma_{(p)}}{\sigma}_m,$$

$$\binom{\sigma_{(p)}^{(m-1,m)}}{\sigma^{(m-1,m)}}_m=j_{\sigma^{(m)}}\left(g^{\sigma^{(m)}}_{\sigma^{(m-1,m)}1}\right)^{-1}j^{-1}_{\sigma_{(p)}^{(m)}}g^{\sigma_{(p)}^{(m)}}_{\sigma_{(p)}^{(m-1,m)}1}\binom{\sigma_{(p)}^{(m)}}{\sigma^{(m)}}_m.$$

Again by (62) the coefficients of

$$\binom{\sigma_{(p)}}{\sigma}_m \quad\text{and}\quad \binom{\sigma_{(p)}^{(m)}}{\sigma^{(m)}}_m$$

are the same. Hence substituting these into (63) yields (61).

The case $d=2$ must be treated separately. Note that $\sigma_{m-1}=\sigma_m$. Using (56), one can easily show that $\binom{\kappa}{\sigma}_m$ is proportional to $\binom{\kappa^{(i)}}{\sigma^{(i)}}_m$ for $i\neq p$, $m-1$, m. Iterating this, for the proof of (61), we may assume that $\sigma_{m-1}=\sigma_{m-2}=\cdots=\sigma_{p+1}$, $\sigma_p+1=\sigma_{p+1}=\cdots=\sigma_1$. Suppose $p\leq m-3$. Then (61) is a consequence of (56) with $\mu=\sigma_{(p)}^{(m)}$ and $\nu=\sigma$. Hence assume $p=m-2$. We shall calculate $\binom{\sigma_{(p)}}{\sigma}_m$ explicitly. It follows from (54) that

$$\begin{aligned}&J_{\sigma_{(p)}}(1+t_1,\dots,1+t_m)\\ &\quad=\prod_{\substack{1\leq i\leq m\\ 1\leq j\leq\sigma_m}}h_*^{\sigma_{(p)}}(i,j)\{(1+t_1)\cdots(1+t_m)\}^{\sigma_m}J_{\sigma_{(p)}-\sigma_m I}(1+t_1,\dots,1+t_m)\\ &\quad=J_{\sigma_{(p)}}(1^m)\prod_{\substack{1\leq i\leq m\\ 1\leq j\leq\sigma_m}}h_*^{\sigma_{(p)}}(i,j)\{(1+t_1)\cdots(1+t_m)\}^{\sigma_m}\sum_\tau\binom{\sigma_{(p)}-\sigma_m I}{\tau}_m\frac{J_\tau(t)}{J_\tau(1^m)},\end{aligned}\tag{64}$$

where $\sigma_{(p)}-\sigma_m I=(\sigma_1-\sigma_m,\dots,\sigma_{m-1}-\sigma_m,0)$. We compute the coefficient of J_σ in the right-hand side of (64). For this it suffices to consider only the terms of $\tau=\sigma-\sigma_m I$ or $\sigma_{(p)}-\sigma_m I$. Since $g^{\lambda}_{\sigma_{(p)}-\sigma_m I,1}m-1\neq 0$ only if $\sigma_{(p)}-\sigma_m I\subseteq\lambda$ and $1^{m-1}\subseteq\lambda$, we see that

$$m_{1^{m-1}}(t)J_{\sigma_{(p)}-\sigma_m I}(t)=c_1t_1\cdots t_mJ_{\sigma-\sigma_m I}(t)+c_2J_{\sigma_{(p)}^{(m)}-(\sigma_m-1)I}(t),$$

which implies that

$$\begin{aligned}\binom{\sigma_{(p)}}{\sigma}_m&=J_\sigma(1^m)J_{\sigma_{(p)}}(1^m)^{-1}\prod_{\substack{1\leq i\leq m\\ 1\leq j\leq\sigma_m}}h_*^{\sigma_{(p)}}(i,j)h_*^{\sigma}(i,j)^{-1}\\ &\quad\cdot\left\{J_{\sigma_{(p)}-\sigma_m I}(1^m)J_{\sigma-\sigma_m I}(1^m)^{-1}\binom{\sigma_{(p)}-\sigma_m I}{\sigma-\sigma_m I}_m+\sigma_m c_1\right\}.\end{aligned}\tag{65}$$

Write $[m_\lambda(t)]f(t)$ as the coefficient of m_λ when the symmetric function $f(t)$ is expanded in terms of the monomial symmetric functions. Put

$$d_1=[t_1\cdots t_m m_{\sigma-\sigma_m I}(t)]m_{1^{m-1}}(t)m_{\sigma_{(p)}-\sigma_m I}(t)$$

$$d_2=[t_1\cdots t_m m_{\sigma-\sigma_m I}(t)]m_{1^{m-1}}(t)m_{\sigma_{(p+1)}-\sigma_m I}(t).$$

Observe that $d_1=1$, $d_2=m-p=2$ (respectively, $d_1=m-p+1=3$, $d_2=0$) when $\sigma_{(p+1)}=\sigma_{(m-1)}$ is admissible, i.e., $\sigma_{m-1}<\sigma_{m-2}$ (respectively, $\sigma_{m-1}=\sigma_{m-2}$). Thus we get

$$\begin{aligned}c_1=v^{-1}_{\sigma-\sigma I_m,\sigma-\sigma_m I}\{d_1 v_{\sigma_{(p)}-\sigma_m I,\sigma_{(p)}-\sigma_m I}+d_2 v_{\sigma_{(p)}-\sigma_m I,\sigma_{(m-1)}-\sigma_m I}\\ -c_2 v_{\sigma^{(m)}_{(p)}-(\sigma_m-1)I,\sigma-(\sigma_m-1)I}\},\end{aligned}\tag{66}$$

in which (55) implies that

$$\begin{aligned}c_2&=v_{\sigma_{(p)}-\sigma_m I,\sigma_{(p)}-\sigma_m I}\cdot v^{-1}_{\sigma^{(m)}_{(p)}-\sigma_m I,\sigma^{(m)}_{(p)}-\sigma_m I}\\ &=\prod_{i=2}^{m-1}(i+\alpha(\sigma_p+1-\sigma_m))^{-1}.\end{aligned}$$

By using (55) and Lemma 8 we have

$$v^{-1}_{\sigma-\sigma_m I,\sigma-\sigma_m I}\cdot v_{\sigma_{(p)}-\sigma_m I,\sigma_{(p)}-\sigma_m I}=p(1+\alpha(\sigma_p-\sigma_m)),$$

$$v^{-1}_{\sigma-\sigma_m I,\sigma-\sigma_m I}\cdot v_{\sigma_{(p)}-\sigma_m I,\sigma_{(m-1)}-\sigma_m I}=p(\sigma_p+1-\sigma_m),$$

$$\begin{aligned}&v^{-1}_{\sigma-\sigma_m I,\sigma-\sigma_m I}\cdot v_{\sigma^{(m)}_{(p)}-(\sigma_m-1)I,\sigma-(\sigma_m-1)I}\\ &=p(\sigma_p+1-\sigma_m)\frac{1+\alpha(\sigma_p+2-\sigma_m)}{1+\alpha(\sigma_p+1-\sigma_m)}(2+\alpha(\sigma_p-\sigma_m))\prod_{i=3}^{m-1}(i+\alpha(\sigma_p+1-\sigma_m))\\ &\quad+p(\sigma_p+2-\sigma_m)\frac{m-1+\alpha(\sigma_p+1-\sigma_m)}{1+\alpha(\sigma_p+1-\sigma_m)}(1+\alpha(\sigma_p-\sigma_m))\prod_{i=2}^{m-2}(i+\alpha(\sigma_p+1-\sigma_m)).\end{aligned}$$

Substituting these to (66) yields (whether $\sigma_{(m-1)}$ is admissible or not)

$$c_1=\alpha p(\sigma_p+1-\sigma_m)(3+\alpha(\sigma_p-\sigma_m))(2+\alpha(\sigma_p+1-\sigma_m))^{-1}.$$

Hence substituting this into (65) and using (15) and Lemma 7, we finally obtain

$$\binom{\sigma_{(p)}}{\sigma}_m=p(\sigma_p+1-\sigma_m)(2+\alpha(\sigma_p+1))(2+\alpha(\sigma_p+1-\sigma_m))^{-1}.$$

This also gives

$$\binom{\sigma^{(m-1,m)}_{(p)}}{\sigma^{(m-1,m)}}_m=p(\sigma_p+2-\sigma_m)(2+\alpha(\sigma_p+1))(2+\alpha(\sigma_p+2-\sigma_m))^{-1}.$$

On the other hand, by virtue of (56), we have

$$\binom{\sigma^{(m)}_{(p)}}{\sigma^{(m)}}_m=j^{-1}_{\sigma^{(m)}}g^{\sigma^{(m)}}_{\sigma^{(m-1,m)}1}j_{\sigma^{(m)}_{(p)}}(g^{\sigma^{(m)}_{(p)}}_{\sigma^{(m-1,m)}_{(p)}1})^{-1}\binom{\sigma^{(m-1,m)}_{(p)}}{\sigma^{(m-1,m)}}_m.$$

Hence it remains only to prove

$$j_{\sigma}^{-1} g^{\sigma}_{\sigma^{(m)}1}\left(j^{-1}_{\sigma_{(p)}} g^{\sigma_{(p)}}_{\sigma^{(m)}_{(p)}1}\right)^{-1} j^{-1}_{\sigma^{(m)}} g^{\sigma^{(m)}}_{\sigma^{(m-1,m)}1}\left(j^{-1}_{\sigma^{(m)}_{(p)}} g^{\sigma^{(m)}_{(p)}}_{\sigma^{(m-1,m)}_{(p)}1}\right)^{-1}$$
$$=(\sigma_p+1-\sigma_m)(2+\alpha(\sigma_p+2-\sigma_m))$$
$$\cdot\{(\sigma_p+2-\sigma_m)(2+\alpha(\sigma_p+1-\sigma_m)\}^{-1},$$

provided $p=m-2$. But this follows at once from (62).

Case $\kappa_{m-1}=\sigma_{m-1}+2$. In this case we see $\kappa_1=\sigma_1,\dots,\kappa_{m-2}=\sigma_{m-2}$. Putting $\mu=\kappa^{(m-1)}$ and $\nu=\sigma$ in (56), one can easily show that $\binom{\kappa}{\sigma}_m=0$ when $\sigma\not\subseteq\kappa$ is equivalent to (61) with $p=m-1$. But the latter can be proved in the same way as in the case of $\kappa_{m-1}=\sigma_{m-1}+1$ using (56). So we omit the details of the proof.

Proof of (b). Induction on $|\kappa|$, the case $|\kappa|=1$ being clear. Suppose $\kappa=\sigma_{(p)}$ and that $\sigma_{(r)}$ are admissible for some $r\neq p$. Put $\mu=\kappa$ and $\nu=\sigma_{(r)}$ in (56). Then by (a) we have

$$\binom{\kappa_{(r)}}{\sigma_{(r)}}_m=\left(j^{-1}_{\kappa_{(r)}}g^{\kappa_{(r)}}_{\kappa 1}\right)^{-1}j^{-1}_{\sigma_{(r)}}g^{\sigma_{(r)}}_{\sigma 1}\binom{\kappa}{\sigma}_m .$$

The coefficient of $\binom{\kappa}{\sigma}_m$ is independent of the dimension m by definition, and the proof follows by induction.

5.3. Expression of $\binom{\kappa}{\sigma}$ by $g^{\lambda}_{\mu\nu}$.

PROPOSITION 2. *We have*

$$\binom{\kappa}{\sigma}=\sum_{\mu}(j_\sigma j_\mu)^{-1}g^{\kappa}_{\sigma\mu}.$$

Proof. Suppose first that $|\kappa|=|\sigma|+1$. Since $\binom{\kappa}{\sigma}$'s are independent of the dimension, we see

$$\begin{aligned}\frac{J_\kappa(1,t_1,\dots,t_m)}{J_\kappa(1^{m+1})}&=\sum_{\lambda\subseteq\kappa}\binom{\kappa}{\lambda}\frac{J_\lambda(t_1-1,\dots,t_m-1)}{J_\lambda(1^{m+1})}\\ &=\sum_{\lambda\subseteq\kappa}\binom{\kappa}{\lambda}J_\lambda(1^{m+1})^{-1}J_\lambda(1^m)\sum_{\sigma\subseteq\lambda}(-1)^{|\lambda|+|\sigma|}\binom{\lambda}{\sigma}\frac{J_\sigma(t)}{J_\sigma(1^m)}.\end{aligned}\tag{67}$$

On the other hand it follows from (53) that

$$J_\kappa(1,t)=\sum_{\substack{\sigma\subseteq\kappa\\ \nu\subseteq\kappa}}g^{\kappa}_{\sigma\nu}J_\nu(1)j_\nu^{-1}J_\sigma(t)j_\sigma^{-1}.\tag{68}$$

As $\binom{\sigma}{\sigma}=1$, equating the coefficients of $J_\sigma(t)$ with $|\sigma|=|\kappa|-1$ in (67) and (68) gives

$$\{J_\kappa(1^{m+1})J_\sigma(1^{m+1})^{-1}-J_\kappa(1^m)J_\sigma(1^m)^{-1}\}\binom{\kappa}{\sigma}=J_1(1)j_1^{-1}j_\sigma^{-1}g^{\kappa}_{\sigma 1}.$$

Clearly $J_1(1)=1$ and (60) imply that the coefficient of $\binom{\kappa}{\sigma}$ is 1, which completes the proof in the case $|\kappa|=|\sigma|+1$.

Suppose now that $|\kappa|-|\sigma|=r$ $(r\geq 2)$ and that the proposition holds in the cases $|\kappa|-|\sigma|<r$. It follows from Lemma 4 (with $d=|\sigma|+1$) that

$$\begin{aligned}r\binom{\kappa}{\sigma}&=\sum_{|\tau|=|\sigma|+1}\binom{\kappa}{\tau}\binom{\tau}{\sigma}\\ &=\sum_{|\tau|=|\sigma|+1}\left(\sum_{|\nu|=r-1}(j_\tau j_\nu)^{-1}g^{\kappa}_{\tau\nu}\right)(j_\sigma j_1)^{-1}g^{\tau}_{\sigma 1}.\end{aligned}$$

Since $j_1=\alpha$, it remains only to prove that

$$\sum_{\tau,\nu}(j_\tau j_\nu)^{-1}g^\kappa_{\tau\nu}g^\tau_{\sigma 1}=\alpha r\sum_\mu j_\mu^{-1}g^\kappa_{\mu\sigma}.$$

Let $m\geq l(\kappa)$. Since $J_1(t)=J_1(t_1,\dots,t_m)=t_1+\cdots+t_m$, it follows from (14) that

$$\begin{aligned}J_\sigma(t)(t_1+\cdots+t_m)^r&=J_\sigma(t)J_1(t)(t_1+\cdots+t_m)^{r-1}\\&=\sum_\tau j_\tau^{-1}g^\tau_{\sigma 1}J_\tau(t)\left(\sum_\nu \alpha^{r-1}(r-1)!\,j_\nu^{-1}J_\nu(t)\right)\\&=\alpha^{r-1}(r-1)!\sum_{\lambda,\nu,\tau}(j_\lambda j_\nu j_\tau)^{-1}g^\tau_{\sigma 1}g^\lambda_{\tau\nu}J_\lambda(t).\end{aligned}\tag{69}$$

On the other hand, we see that

$$\begin{aligned}J_\sigma(t)(t_1+\cdots+t_m)^r&=J_\sigma(t)\left(\sum_\mu \alpha^r r!\,j_\mu^{-1}J_\mu(t)\right)\\&=\alpha^r r!\sum_{\lambda,\mu}(j_\lambda j_\mu)^{-1}g^\lambda_{\mu\sigma}J_\lambda(t).\end{aligned}\tag{70}$$

Comparing the coefficients of $J_\kappa(t)$ in (69) and (70) completes the proof.

Remark. Theorem 3 and Proposition 2 have been announced also by Lassalle [L]. The full proofs have not yet appeared as of July 1992. Our proof of Theorem 3 is a rather messy case-by-case check. It would be interesting to find an intrinsic proof. (Of course, it will be better to give a direct proof of Proposition 2.)

6. Consequences.

6.1. Main result. The following transformation formula of solutions of the system (9) can be proved in the exact same way as in the case of the Gaussian hypergeometric equation and we omit the proof.

PROPOSITION 3. *If $F(t_1,\dots,t_m)$ is a solution of the system* (9), *then $(t_1\cdots t_m)^{-a}F(1/t_1,\dots,1/t_m)$ is also a solution of the system obtained from* (9) *by replacing b by $a-c+1+(1/\alpha)(m-1)$ and c by $a-b+1+(1/\alpha)(m-1)$.*

Note that the integral

$$\begin{aligned}&(-1)^{\lambda mn/2}(t_1\cdots t_m)^{-\lambda n/2}S_{n,m}\left(\lambda_1,\lambda_2,\lambda,\frac{-\lambda}{2};\frac{1}{t_1},\dots,\frac{1}{t_m}\right)\\&\quad=\int_{[0,1]^n}\prod_{i=1}^n x_i^{\lambda_1}(1-x_i)^{\lambda_2}\prod_{1\leq i<j\leq n}|x_i-x_j|^\lambda\prod_{\substack{1\leq i\leq n\\1\leq k\leq m}}(1-x_it_k)^{-\lambda/2}\,dx_1\cdots dx_n\end{aligned}\tag{71}$$

is analytic at the origin. Combining Theorems 1, 4, and Proposition 3 gives our main result in the following.

THEOREM 5. *We have*

$$\begin{aligned}&\int_{[0,1]^n}\prod_{\substack{1\leq i\leq n\\1\leq k\leq m}}(x_i-t_k)D_{\lambda_1,\lambda_2,\lambda}(x_1,\dots,x_n)\,dx_1\cdots dx_n\\&\quad=C_1\cdot{}_2F_1^{\lambda/2}\left(-n,\frac{2}{\lambda}(\lambda_1+\lambda_2+m+1)+n-1;\frac{2}{\lambda}(\lambda_1+m);t_1,\dots,t_m\right),\end{aligned}\tag{72}$$

$$\begin{aligned}&\int_{[0,1]^n} \prod_{\substack{1\le i\le n\\ 1\le k\le m}} (1-x_it_k)^{-\lambda/2} D_{\lambda_1,\lambda_2,\lambda}(x_1,\dots,x_n)\,dx_1\cdots dx_n\\ &\quad= C_2\cdot {}_2F_1^{2/\lambda}\left(\frac{\lambda}{2}n, \frac{\lambda}{2}(n-1)+\lambda_1+1; \lambda(n-1)+\lambda_1+\lambda_2+2; t_1,\dots,t_m\right),\end{aligned}\tag{73}$$

where $C_1=S_{n,0}(\lambda_1+m, \lambda_2, \lambda)$ *and* $C_2=S_{n,0}(\lambda_1, \lambda_2, \lambda)$.

This theorem gives a rather short proof of Macdonald's conjecture (C5) [Ma3, p. 197]. Namely, the following.

COROLLARY 2. *Let* $\kappa=(\kappa_1, \kappa_2, \dots)$ *be a partition of length* $\le n$. *Then*

$$\begin{aligned}&\int_{[0,1]^n} J_\kappa^{2/\lambda}(x_1,\dots,x_n)D_{\lambda_1,\lambda_2,\lambda}(x_1,\dots,x_n)\,dx_1\cdots dx_n\\ &\quad=J_\kappa^{2/\lambda}(1,\dots,1)\\ &\qquad\cdot\prod_{i=1}^{n}\frac{\Gamma(i(\lambda/2)+1)\Gamma(\kappa_i+\lambda_1+(\lambda/2)(n-i)+1)\Gamma(\lambda_2+(\lambda/2)(n-i)+1)}{\Gamma((\lambda/2)+1)\Gamma(\kappa_i+\lambda_1+\lambda_2+(\lambda/2)(2n-i-1)+2)}.\end{aligned}\tag{74}$$

Proof. Let $m\ge n$. We notice the Cauchy identity [St, Prop. 2.1, p. 79]:

$$\sum_\nu J_\nu^{(\alpha)}(x_1,\dots,x_n)J_\nu^{(\alpha)}(t_1,\dots,t_m)j_\nu^{-1}=\prod_{\substack{1\le i\le n\\ 1\le k\le m}}(1-x_it_k)^{-1/\alpha},\tag{75}$$

where the summation is over all partitions ν. Put $\alpha=2/\lambda$ and substitute the left-hand side of (75) into that of (73). Equating the coefficients of $J_\kappa^{(2/\lambda)}(t_1,\dots,t_m)$ of both sides gives

$$\begin{aligned}&\int_{[0,1]^n} J_\kappa^{2/\lambda}(x_1,\dots,x_n)D_{\lambda_1,\lambda_2,\lambda}(x_1,\dots,x_n)\,dx_1\cdots dx_n\\ &\quad=C_2\cdot\frac{[(\lambda/2)n]_\kappa^{(\alpha)}[(\lambda/2)(n-1)+\lambda_1+1]_\kappa^{(\alpha)}}{[\lambda(n-1)+\lambda_1+\lambda_2+2]_\kappa^{(\alpha)}}\left(\frac{2}{\lambda}\right)^{|\kappa|}.\end{aligned}$$

Now the desired equality (75) follows at once from (1), (15), (18), and (19).

5.2. Case of $m=1,2$. In case $m=1$, putting $a=(\lambda/2)n$, $b=(\lambda/2)(n-1)+\lambda_1+1$, $c=\lambda(n-1)+\lambda_1+\lambda_2+2$ in (73) yields the following integral representation of the Gauss hypergeometric function:

$$\begin{aligned}{}_2F_1(a, b; c; t)=C_3\cdot&\int_{[0,1]^n}\prod_{i=1}^{n}x_i^{b-a+(a/n)-1}(1-x_i)^{c-a-b+(a/n)-1}\prod_{1\le i<j\le n}|x_i-x_j|^{2a/n}\\ &\cdot\prod_{i=1}^{n}(1-x_it)^{-a/n}\,dx_1\cdots dx_n\\ C_3=S_{n,0}&\left(b-a+\frac{a}{n}-1, c-a-b+\frac{a}{n}-1, \frac{2a}{n}\right)^{-1}.\end{aligned}$$

In case $m=2$, one can easily deduce from (10) that

$$\begin{aligned}{}_2F_1^{(\alpha)}(a, b; c; t_1, t_2)=C_4\cdot F_4&\left(a, b, c-\frac{1}{\alpha}, a+b-c+1, t_1t_2, (1-t_1)(1-t_2)\right)\\ &+C_5\cdot((1-t_1)(1-t_2))^{c-a-b}\\ &\cdot F_4\left(c-b, c-a, c-\frac{1}{\alpha}, c-a-b+1, t_1t_2, (1-t_1)(1-t_2)\right).\end{aligned}\tag{76}$$

Putting $t_2=0$ in (76) yields

$$\begin{aligned}{}_2F_1(a,b;c;t_1)&=C_4\cdot{}_2F_1(a,b;a+b-c+1;1-t_1)\\&\quad+C_5(1-t_1)^{c-a-b}\,{}_2F_1(c-b,c-a;c-a-b+1;1-t_1).\end{aligned}$$

The connection formula of solutions of the Gauss hypergeometric equation [W-W, p. 291] then gives the values of C_4 and C_5:

$$C_4=\frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)},\qquad C_5=\frac{\Gamma(c)\Gamma(a+b-c)}{\Gamma(a)\Gamma(b)}.$$

Combining (73) and (76), we get

$$\begin{aligned}&\int_{[0,1]^n}\prod_{\substack{1\le i\le n\\k=1,2}}(1-x_it_k)^{-\lambda/2}D_{\lambda_1,\lambda_2,\lambda}(x_1,\dots,x_n)\,dx_1\cdots dx_n\\&=C_2C_4\cdot F_4\Big(\frac{\lambda}{2}n,\frac{\lambda}{2}(n-1)+\lambda_1+1,\frac{\lambda}{2}(2n-3)+\lambda_1+\lambda_2+2,\frac{\lambda}{2}-\lambda_2;t_1t_2,(1-t_1)(1-t_2)\Big)\\&\quad+C_2C_5((1-t_1)(1-t_2))^{\lambda_2-(\lambda/2)+1}F_4\Big(\frac{\lambda}{2}(n-1)+\lambda_2+1,\frac{\lambda}{2}(n-2)+\lambda_1+\lambda_2+2,\\&\qquad\qquad\frac{\lambda}{2}(2n-3)+\lambda_1+\lambda_2+2,\\&\qquad\qquad\lambda_2-\frac{\lambda}{2}+2;t_1t_2,(1-t_1)(1-t_2)\Big).\end{aligned}$$

7. Relation with generalized Jacobi polynomials.

For $\alpha>-1$, $\beta>-1$, and $\gamma\ge-\frac{1}{2}$ put

$$w(x)=w^{\alpha,\beta,\gamma}(x)=\prod_{i=1}^m(1-x_i)^\alpha(1+x_i)^\beta\prod_{1\le i<j\le m}(x_i-x_j)^{2\gamma+1},\qquad x\in\Omega$$

where Ω is the region

$$\Omega=\{x\in\mathbb{R}^m\,|\,-1\le x_m\le x_{m-1}\le\cdots\le x_1\le 1\}.$$

As a total ordering of partitions we still use the reverse lexicographic order $\le^R$: For partitions κ and ν, denote $\nu\le^R\kappa$ if $\nu=\kappa$ or the first nonvanishing difference $\kappa_i-\nu_i$ is positive (we do not assume $|\kappa|=|\nu|$). Hereafter we will restrict partitions to have length $\le m$. A symmetric polynomial $p(x)$ is said to have degree κ with leading term $c_\kappa m_\kappa$ if

$$p(x)=\sum_{\nu\le^R\kappa}c_\nu m_\nu,\qquad c_\kappa\ne 0.$$

Definition [Ko], [V], [D]. The *generalized Jacobi polynomials* $p_\kappa^{\alpha,\beta,\gamma}(x)$ are defined by

(1) $p_{(0)}=1$;

(2) $p_\kappa^{\alpha,\beta,\gamma}(x)$ is a symmetric polynomial with leading term m_κ;

(3) $\int_\Omega p_\kappa^{\alpha,\beta,\gamma}(x)q(x)w^{\alpha,\beta,\gamma}(x)\,dx=0$ if $q(x)$ is a symmetric polynomial and degree $q\le^R\kappa$.

THEOREM 6. *We have*

$$S_{n,m}(\lambda_1,\lambda_2,\lambda,1;t_1,\dots,t_m)=2^{-n}\cdot S_{n,0}(\lambda_1,\lambda_2,\lambda)p^{\alpha,\beta;\gamma}_{(n^m)}(1-2t_1,\dots,1-2t_m),$$

where $\alpha=(2/\lambda)(\lambda_1+1)-1$, $\beta=(2/\lambda)(\lambda_2+1)-1$, $\gamma=(\lambda/2)-(1/2)$, *and* (n^m) *is the partition* $(n,\dots,n)$.

Proof. Note first that the leading term of

$$S_{n,m}(\lambda_1,\lambda_2,\lambda,1;(1-t_1)/2,\dots,(1-t_m)/2) \text{ is } 2^{-n}S_{n,0}(\lambda_1,\lambda_2,\lambda)(t_1\cdots t_m)^n.$$

Hence it remains to prove

$$\int_{[0,1]^m} S_{n,m}(\lambda_1,\lambda_2,\lambda;t)q(t)D_{\alpha,\beta,2\gamma+1}(t)\,dt=0, \tag{77}$$

where $q(t)$ is a symmetric polynomial with degree $<^R(n^m)$. Clearly it will be sufficient to consider the case $q(t)=J^{(\lambda/2)}_\kappa(t)$ with $\kappa<^R(n^m)$. Put

$$\Delta^{(\alpha,\beta,\gamma)}=\sum_{i=1}^m\left\{t_i(1-t_i)\frac{\partial^2}{\partial t_i^2}+\left[\alpha+1-(\alpha+\beta+2)t_i+(2\gamma+1)\sum_{\substack{j=1\\ j\neq i}}^m\frac{t_i(1-t_i)}{t_i-t_j}\right]\frac{\partial}{\partial t_i}\right\}.$$

Then for polynomials f and g we have [V, Thm. 4.3]

$$\int_{[0,1]^m}(\Delta^{(\alpha,\beta,\gamma)}f)gD_{\alpha,\beta,2\gamma+1}(t)\,dt=\int_{[0,1]^n}f(\Delta^{(\alpha,\beta,\gamma)}g)D_{\alpha,\beta,2\gamma+1}(t)\,dt. \tag{78}$$

One can easily deduce from (44) that

$$\Delta^{(\alpha,\beta,\gamma)}S_{n,m}(\lambda_1,\lambda_2,\lambda,1;t)=-\frac{2mn}{\lambda}\left(\lambda_1+\lambda_2+m+1+\frac{\lambda}{2}(n-1)\right)S_{n,m}(\lambda_1,\lambda_2,\lambda,1;t).$$

Now suppose that (77) holds for any symmetric polynomial with degree $<^R\kappa$ and that $\kappa<^R(n^m)$. We prove

$$\int_{[0,1]^n}S_{n,m}(\lambda_1,\lambda_2,\lambda,1;t)J^{\lambda/2}_\kappa(t)D_{\alpha,\beta,2\gamma+1}(t)\,dt=0. \tag{79}$$

It is readily verified by using (11), (12), and (37) that

$$\Delta^{(\alpha,\beta,\gamma)}J^{\lambda/2}_\kappa(t)=-\frac{2}{\lambda}\left(2e_\kappa\left(\frac{\lambda}{2}\right)+(\lambda_1+\lambda_2+2)|\kappa|\right)J^{\lambda/2}_\kappa(t)+q(t),$$

where $q(t)$ is a symmetric polynomial with degree $<^R\kappa$. On using (13), we observe that

$$\frac{2mn}{\lambda}\left(\lambda_1+\lambda_2+m+1+\frac{\lambda}{2}(n-1)\right)\neq\frac{2}{\lambda}\left(2e_\kappa\left(\frac{\lambda}{2}\right)+(\lambda_1+\lambda_2+2)|\kappa|\right)$$

provided $\lambda>0$ and $\lambda_1+\lambda_2>0$. Hence (79) follows from (78) for such λ, λ_1, and λ_2. For other values of λ_1 and λ_2, (79) holds by analytic continuation. This completes the proof of the theorem.

REFERENCES

[Ao1] K. AOMOTO, *Jacobi polynomials associated with Selberg integrals*, SIAM J. Math. Anal., 18 (1987), pp. 545-549.

[Ao2] ———, *Gauss-Manin connection of integral of difference products*, J. Math. Soc. Japan, 39 (1987), pp. 191-208.

[Ao3] ———, *Correlation Functions of the Selberg Integral*, in Ramanujan Revisited, G. E. Andrews et al., eds., Academic Press, Boston, MA, 1988, pp. 591-605.

[As] R. ASKEY, *Some basic hypergeometric extensions of integrals of Selberg and Andrews*, SIAM J. Math. Anal., 11 (1980), pp. 938-951.

[A-K] P. APPELL AND J. KAMPÉ DE FÉRIET, *Fonctions hypergéométriques et hypersphériques-polynômes d'Hermite*, Gauthier-Villars, Paris, 1926.

[C] A. G. CONSTANTINE, *Some noncentral distribution problems in multivariate analysis*, Ann. Math. Statist., 34 (1963), pp. 1270-1285.

[D] A. DEBIARD, *Système différentiel hypergéométrique et parties radiales des opérateurs invariants des espaces symétriques de type BC_p*, in Séminaire d'algèbre, M.-P. Malliavin, ed., Lecture Notes in Math. Vol. 1293, Springer, Berlin, 1988, pp. 42-124.

[H] C. S. HERZ, *Bessel functions of matrix argument*, Ann. Math., 61 (1955), pp. 474-523.

[J] A. T. JAMES, *A generating function for averages over the orthogonal group*, Proc. Roy. Soc. London, A229 (1955), pp. 367-375.

[Kad] K. W. J. KADELL, *The Selberg-Jack polynomials*, preprint.

[Kat] M. KATO, *A pfaffian system of Appell's F_4*, Bull. Coll. Ed. Univ. Ryukyus, 33 (1988), pp. 331-334.

[Ko] T. H. KOORNWINDER, *Orthogonal polynomials in two variables which are eigenfunctions of two algebraically independent partial differential operators.* I, II, Indag. Math., 36 (1974), pp. 48-66.

[Kor] A. KORÁNYI, *Hua-type integrals, hypergeometric functions and symmetric polynomials*, in Proceeding of a Conference in Memory of L. K. Hua, Beijing, 1988, to appear.

[L] M. LASSALLE, *Une formule du binôme généralisée pour les polynômes de Jack*, C.R. Acad. Sci. Paris Sér. I Math., 310 (1990), pp. 253-256.

[Ma1] I. G. MACDONALD, *Symmetric Functions and Hall Polynomials*, Oxford University Press, Oxford, 1979.

[Ma2] ———, *Some conjectures for root systems*, SIAM J. Math. Anal., 13 (1982), pp. 988-1007.

[Ma3] ———, *Commuting differential operators and zonal spherical functions*, in Algebraic Groups, Utrecht, 1986, A. M. Cohen et al., eds., Lecture Notes in Math. Vol. 1271, Springer, Berlin, 1987, pp. 189-200.

[Mu] R. J. MUIRHEAD, *Systems of partial differential equations for hypergeometric functions of matrix argument*, Ann. Math. Statist., 41 (1970), pp. 991-1001.

[R] A. REGEV, *Asymptotic values for degrees associated with strips of Young diagrams*, Adv. Math., 41 (1981), pp. 115-136.

[Se] A. SELBERG, *Bemerkninger om et multipelt integral*, Norsk Mat. Tidsskr., 26 (1944), pp. 71-78.

[St] R. P. STANLEY, *Some combinatorial properties of Jack symmetric functions*, Adv. Math., 77 (1989), pp. 76-115.

[V] L. VRETARE, *Formulas for elementary spherical functions and generalized Jacobi polynomials*, SIAM J. Math. Anal., 15 (1984), pp. 805-833.

[W-W] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Cambridge University Press, London, 1927.

[Y] Z. YAN, *Generalized hypergeometric functions*, C.R. Acad. Sci. Paris Sér. I Math., 310 (1990), pp. 349-354.

SIAM J. MATH. ANAL.
Vol. 24, No. 5, pp. 1111–1144, September 1993

001

THE INITIAL VALUE PROBLEM FOR A SYSTEM MODELLING UNIDIRECTIONAL LONGITUDINAL ELASTIC-PLASTIC WAVES*

MICHAEL SHEARER† AND DAVID G. SCHAEFFER‡

Abstract. The authors analyze initial value problems for a hyperbolic system of equations that is a simplification of models of dynamic longitudinal elastoplastic deformations in a rod with hardening. The simplified system has a positive characteristic speed associated with stress waves, and a zero speed associated with the time independence of hardening during elastic deformation. The equations are piecewise linear in stress derivatives, and thus fully nonlinear. The main result is that for bounded uniformly continuous initial data, the Cauchy problem has a unique continuous solution that can be approximated by piecewise linear solutions of the equations.

Key words. elastoplasticity, hyperbolic equations, fully nonlinear partial differential equations, initial value problem, weak solutions

AMS subject classifications. 35L45, 35L60 73E50, 73K05

1. Introduction. In this paper we analyze initial value problems for a hyperbolic system of equations that is a simplification of models of longitudinal elastoplastic deformations in a rod. In longitudinal deformation there is only one nontrivial component of stress, which we label σ, and one component of velocity, w. We consider models with hardening, in which the yield stress (the threshold beyond which the material deforms plastically) depends on the stress history. This dependence is incorporated into the hardening variable γ, which is the maximum previous stress at a material point. The longitudinal elastoplasticity models, studied by many authors [1], [4], [6], [7], are given in [5] in the form

$$(1.1)\qquad \begin{array}{ll} \text{(a)} & \partial_t w = \partial_x \sigma, \\ \text{(b)} & \partial_t \sigma + K\partial_t \gamma = \partial_x w, \\ \text{(c)} & \partial_t \gamma = \begin{cases} 0 & \text{if} \quad \sigma < \gamma \quad \text{(elastic)}, \\ (\partial_t \sigma)_+ & \text{if} \quad \sigma = \gamma \quad (\text{plastic if } \partial_t \sigma \geq 0). \end{cases} \end{array}$$

Here, $K = K(\gamma) > -1$ is a given nonzero function, ∂_t, ∂_x are contractions of $\partial/\partial t, \partial/\partial x$, and

$$(1.2)\qquad (\partial_t \sigma)_+ = \begin{cases} \partial_t \sigma & \text{if } \partial_t \sigma \geq 0, \\ 0 & \text{if } \partial_t \sigma \leq 0. \end{cases}$$

System (1.1) has elastic wave speeds ±1 and 0, and plastic wave speeds $\pm(1+K)^{-1/2}$.

In [5], we simplified the model by taking $K(\gamma)$ constant. Here, we not only take K to be constant, but we also reduce the dimension of the system, eliminating one characteristic variable, corresponding to waves with negative speed. Specifically, we

* Received by the editors March 2, 1992; accepted for publication (in revised form) December 10, 1992.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695. This research was supported by Army Research Office grant DAAL03-91-G-0122 and by National Science Foundation grants DMS-8907736 and DMS-9201115.

‡ Department of Mathematics, Duke University, Durham, North Carolina 27706. The research of this author was supported by National Science Foundation grant DMS-8804592, which includes funds from the Air Force Office of Scientific Research.

write a system of two equations relating stress and hardening, dropping the velocity:

$$\text{(1.3)}\qquad \begin{aligned} &\text{(a)}\quad \partial_t u + k\partial_t v + \partial_x u = 0,\\ &\text{(b)}\quad \partial_t v = \begin{cases} 0 & \text{if } \ u < v \quad \text{(elastic)},\\ (\partial_t u)_+ & \text{if } \ u = v \quad \text{(plastic if } \partial_t u \geq 0\text{)}. \end{cases} \end{aligned}$$

For simplicity, we have relabelled the stress as u, and the hardening as v. The parameter $k > -1$ is related to K by $k = (1+K)^{1/2} - 1$. We call system (1.3) the *unidirectional model.*

The material deforms *elastically* when $u < v$ or $u = v$ and $\partial_t u \leq 0$. Then (1.3) reduces to the system

$$\text{(1.4)}\qquad \begin{aligned} \partial_t u + \partial_x u &= 0,\\ \partial_t v &= 0, \end{aligned}$$

which has wave speeds $c_o = 0$ and $c_E = 1$. We have normalized the elastic wave speed c_E to be one. Note that the variable v is constant in time in elastic deformation.

When $u = v$ and $\partial_t u \geq 0$, the material deforms *plastically*, and (1.3) reduces to a scalar equation:

$$\text{(1.5)}\qquad \partial_t u + \frac{1}{1+k}\partial_x u = 0,$$

with (plastic) wave speed $c_P = (1+k)^{-1}$. By imposing the additional restriction $k > -1$, we ensure that both elastic and plastic waves have positive wave speed. Note that for $-1 < k < 0$, the plastic wave speed is greater than the elastic wave speed, whereas for $k > 0$, the elastic wave speed is larger. For longitudinal motion, the range $-1 < k < 0$ is, therefore, unphysical. However, in reducing the equations for multidimensional deformation with a nonassociative flow rule to a one-dimensional model, the entire range $k > -1$ assumes physical significance. This point is discussed further in [5].

Since solutions of elastoplasticity problems appear to have no spontaneous tendency to form shocks in finite time, we shall be concerned only with continuous solutions in this paper. In particular, for system (1.3) with k constant, the equations are piecewise linear, so that the only shocks expected in solutions of initial boundary value problems are those propagating from initial or boundary conditions. The restriction to continuous solutions also helps to justify the simplification of taking k to be constant. The objective in studying the simplified system is to understand the behavior of solutions locally in space and time, so that the solution is close to a constant, making k close to constant. The principal nonlinearity involved is, therefore, the switch between (1.4) and (1.5). Dropping the quasilinear dependence of the equations through the function k enables us to focus on the nonlinearity due to the switch.

To see how solutions of (1.1) are related to those of (1.3), consider a solution of (1.1) which is unidirectional in the sense that w, σ in each elastic or plastic region are functions of $x - ct$, with $c = 1$ (for elastic deformation), or $c = (1+K)^{-1/2}$ (for plastic deformation). Correspondingly, $\partial_t \gamma = 0$, or $\gamma = \sigma$ (respectively). Then in particular, we have the equation

$$c^{-1}\partial_t \sigma + \partial_x \sigma = 0.$$

This equation may be written in the form (1.3), setting $u = \sigma, v = \gamma, k = (1+K)^{1/2} - 1$. Conversely, if (σ, γ) is a piecewise smooth solution of (1.3), then σ is a function of $x - ct$

in each elastic or plastic region. If w is defined in an elastic or plastic region up to a constant by (1.1a,b), then the triple (w, σ, γ) satisfies (1.1) in that region. However, solutions of (1.1) typically are not unidirectional, and therefore do not satisfy (1.3). Moreover, it is generally not possible to construct the additional variable w in (1.1) so that a solution of (1.3) corresponds to a unidirectional solution of (1.1).

The purpose of the simplified model is to isolate the role of fully nonlinear wave interactions in the global behavior of solutions, while retaining sufficient structure in the equations to preserve the main phenomena. This paper may be viewed as a step toward proving existence results for initial boundary value problems for the full equations (1.1). We show some of the analytical difficulties inherent in dynamic elastoplasticity models, and demonstrate how they are resolved for the simplified system. The additional complication of having waves propagating in both directions (left and right) is an obstacle to extending the results of this paper to the larger system (1.1), because of the possibility of sustained wave interactions through repeated reflections.

The main result is that for bounded uniformly continuous initial data, the Cauchy problem for system (1.3) has a unique solution within a suitable class of bounded continuous functions. The proof of this result is based in part upon ideas related to those of Glimm in his treatment of the Cauchy problem for systems of hyperbolic conservation laws [2]. In particular, we approximate the solution by continuous piecewise linear solutions. Discontinuities in the first derivatives of the approximate solutions are treated as waves that propagate and interact. We use a functional of the piecewise linear approximate solutions that bounds the total spatial variation, is decreasing in time, and estimates the number of discontinuities. To show that this functional is decreasing, we rely heavily on the unidirectional property of the simplified system (1.3).

It is instructive to consider the evolution of initial data shown in Fig. 1.1(a). Initially, $u < v$ everywhere, so that according to (1.4), the graph of v remains fixed while the graph of u moves to the right with speed $c_E = 1$. At some time $t = t_1$, the graph of u first touches the graph of v. At this moment plastic deformation begins and the graph of v is pushed up where it is in contact with the graph of u (i.e., the maximum stress $v(x,t)$ increases). The subsequent motion is shown as a succession of pictures in Fig. 1.1. (The times t_i shown in Fig. 1.1 are defined in Fig. 1.2.) Figure 1.2 shows the (x,t)-plane for this solution, in which a region of plastic deformation (where (1.5) is satisfied) is surrounded by a region of elastic deformation (where (1.4) is satisfied). The solution shown in Fig. 1.1 may be obtained using the method of characteristics, but it is not clear how to use the method of characteristics to prove a general existence result for the Cauchy problem because the boundary between regions of elastic and plastic deformation must be found as part of the solution.

Our analysis of continuous solutions is based upon a detailed treatment of the *scale invariant initial value problem* (SI problem for short), which is the initial value problem for (1.3) with piecewise linear initial conditions:

$$(1.6)\qquad \begin{aligned} &\text{(a)} \qquad u(x,0) = \begin{cases} a_L x & \text{if} \quad x < 0, \\ a_R x & \text{if} \quad x > 0. \end{cases} \\ &\text{(b)} \qquad v(x,0) = \begin{cases} b_L x & \text{if} \quad x < 0, \\ b_R x & \text{if} \quad x > 0. \end{cases} \end{aligned}$$

To ensure that $u \leq v$ everywhere, we require that

$$(1.7)\qquad a_L \geq b_L, \qquad a_R \leq b_R.$$

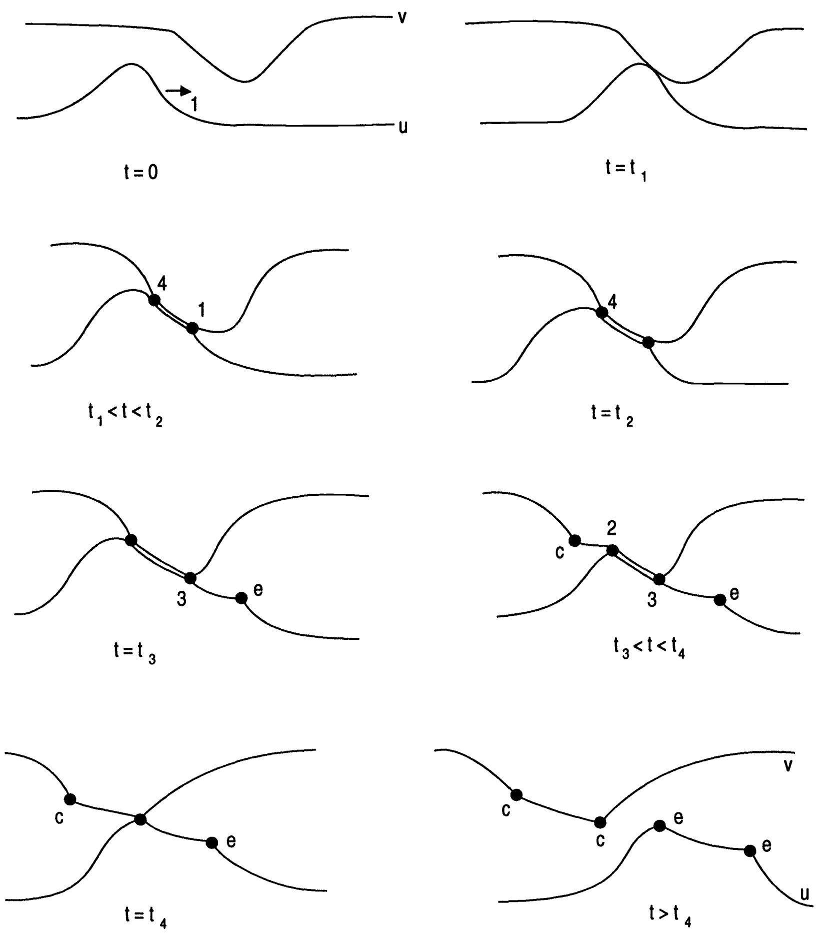


FIG. 1.1. *Labels refer to the classification of waves in* §2.

The initial value problem (1.3), (1.6) is called scale invariant because (1.3) and the initial conditions (1.6) are unchanged by the scaling

$$\begin{aligned} x &\longmapsto \alpha x, & t &\longmapsto \alpha t, \\ u &\longmapsto \alpha^{-1} u, & v &\longmapsto \alpha^{-1} v \end{aligned} \tag{1.8}$$

for any $\alpha > 0$.

In a companion paper [5], we solve the corresponding SI problem for the larger system (1.1). In this paper, we prove the existence of a solution of the Cauchy problem, with general initial data, but only for the simpler system of (1.3). In the proof, we use detailed information about solutions of SI problems for (1.3). We treat the SI problem for (1.3) in §2 for $k > 0$. The solution is continuous and piecewise linear. Jumps in derivatives of the solution propagate along characteristics, or along boundaries between regions of elastic and plastic deformation, which we call fronts or elastic-plastic boundaries. In §2, we classify the fronts into four types. An interesting point here is that one of the types of front has negative speed, even though all characteristic speeds are nonnegative.

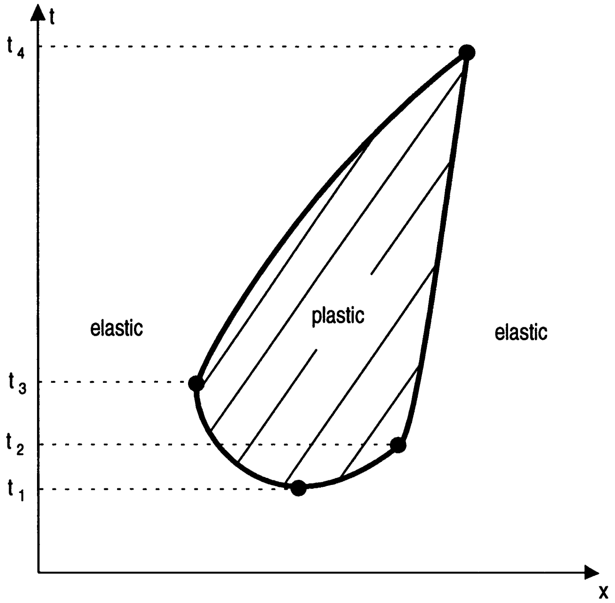


FIG. 1.2. *Elastic and plastic regions.*

In §3, we solve the Cauchy problem for bounded continuous initial data. The solution is continuous globally in time, and it is unique within the class of solutions we seek, namely those that can be approximated by continuous piecewise linear solutions of initial value problems having piecewise linear initial conditions.

In Appendix A, we consider the SI problem when $k < 0$ in (1.3). We demonstrate nonuniqueness of solutions of the SI problem for certain initial data, and nonexistence of continuous solutions for other initial data. Nonuniqueness can be resolved by imposing an entropy condition analogous to the Lax entropy condition for shock waves. We conjecture that nonexistence may be overcome by introducing a class of discontinuous solutions.

2. Solution of the scale invariant problem for $k > 0$.

For a given solution of (1.3), the (x, t)-plane is divided into regions in which either (1.4) is satisfied or (1.5) is satisfied. We refer to these regions as *elastic* regions or *plastic* regions, respectively. Within an elastic region there are characteristics with speed $c_E = 1$, which we call *elastic* characteristics, and characteristics with speed $c_o = 0$, which we call *contact* characteristics. Within a plastic region, there is just one family of characteristics, the *plastic* characteristics, which have speed $c_P = (1 + k)^{-1}$.

We seek a continuous piecewise differentiable solution of the SI problem (1.3), (1.6) when $k > 0$. If the solution is unique, then it must be invariant under the scaling (1.8). Therefore, we suppose that

$$(u,v)(x,t) = t\left(f\left(\frac{x}{t}\right), g\left(\frac{x}{t}\right)\right) \tag{2.1}$$

for some functions f and g. For such a solution, the conditions for plastic or elastic deformation are invariant under the scalings $x \to \alpha x, t \to \alpha t$ for any $\alpha > 0$. Therefore the elastic and plastic regions for a solution of the SI problem form wedges in the (x, t)−plane with vertices at the origin. The following lemma establishes the property that solutions of the SI problem are piecewise linear.

LEMMA 2.1. *Within a plastic region not containing the plastic characteristic $x = c_P t$, or within an elastic region not containing the elastic characteristic $x = c_E t$ or the contact characteristic $x = 0$, a solution of* (1.3) *of the form* (2.1) *is linear.*

Proof. In a plastic region, $u = v$, so that $f = g$ in (2.1). Substituting into (1.5) we obtain

$$f - \xi f' + (1+k)^{-1} f' = 0, \tag{2.2}$$

where $\xi = x/t$. By hypothesis, $\xi \neq c_P = (1+k)^{-1}$. Therefore, (2.2) has the general solution $f(\xi) = a(\xi - (1+k)^{-1})$, where a is an arbitrary constant. In this case,

$$u(x,t) = v(x,t) = ax - a(1+k)^{-1} t. \tag{2.3}$$

In an elastic region, substituting (2.1) into (1.4) gives the pair of equations

$$\begin{aligned} f - \xi f' + f' &= 0, \\ g - \xi g' &= 0. \end{aligned} \tag{2.4}$$

Since $\xi \neq 1$ and $\xi \neq 0$ by hypothesis, we obtain

$$f(\xi) = a(\xi - 1), \qquad g(\xi) = b\xi,$$

for which

$$u(x,t) = a(x-t), \qquad v(x,t) = bx. \tag{2.5}$$

This completes the proof. □

Note that solution (2.3) in a plastic region has one free parameter a, whereas solution (2.5) in an elastic region has two free parameters, a and b. The structure of solutions of the SI problem is a fan of adjacent wedges in the (x,t)-plane. In each wedge the solution is linear and specified by a point (a,b) in the plane, with $a = b$ if the wedge lies in a plastic region. The boundaries $x/t =$ constant between adjacent wedges are rays across which the solution is continuous but its derivative has a jump corresponding to a jump in the pair (a,b). We call such a ray a *single wave*. Therefore, we may represent the solution of the SI problem as a sequence of points in the (a,b) plane. To start with, we classify all the different types of single waves.

2.1. Single waves. Consider a single wave $x/t = s$ with speed s across which (u,v) is continuous but the first derivative of (u,v) may jump. We use the notation (u_E, v_E) for (u,v) if the material is deforming elastically, and the notation (u_P, v_P) at points of plastic deformation. For a function $w(x,t)$, the notation $[w]$ means the jump in w across the wave:

$$[w] = w(st+, t) - w(st-, t).$$

From Lemma 2.1, we know that within an elastic region or a plastic region, the only possible single waves are characteristics. We classify these in the usual way for hyperbolic equations, giving the corresponding changes across the waves in the parameters a and b, which are the spatial derivatives of the variables u, v. The time derivatives are determined from the differential equations by the spatial derivatives.

Elastic waves. Elastic on both sides.

$$[a] \neq 0, \quad [b] = 0, \quad s = 1.$$

Contact waves. Elastic on both sides.

$$[a] = 0, \quad [b] \neq 0, \quad s = 0.$$

Plastic waves. Plastic on both sides.

$$[a] = [b] \neq 0, \qquad s = (1+k)^{-1}.$$

Elastic-plastic boundaries. These are single waves $x = st$. On one side the material deforms elastically, and on the other side the material deforms plastically. Compatibility conditions for these waves are derived [4] using continuity of the functions u, v across the wave, and the appropriate differential equation from (1.3) on each side of the boundary.

Continuity at $x = st$ yields the conditions

$$\begin{aligned} u_E(st,t) &= u_P(st,t), \\ v_E(st,t) &= v_P(st,t) = u_P(st,t). \end{aligned}$$

We use one equation to discard v_P as an unknown, and differentiate the remaining equations, recalling $\partial_t v_E = 0$, to obtain

$$\begin{aligned} (\partial_t + s\partial_x)u_E &= (\partial_t + s\partial_x)u_P, \\ s\partial_x v_E &= (\partial_t + s\partial_x)u_P, \end{aligned} \tag{2.6}$$

From (1.3),

$$\begin{aligned} \partial_t u_E &= -\partial_x u_E, \\ \partial_t u_P &= -(1+k)^{-1}\partial_x u_P. \end{aligned} \tag{2.7}$$

Eliminating $\partial_x u_P$ and $\partial_x u_E$ from (2.6), (2.7), we get

$$\frac{\partial_t u_E}{\partial_t u_P} = \frac{1 - s(1+k)}{1-s}. \tag{2.8}$$

For future use, we record the result of solving (2.6), (2.7) for s and for $a_P = \partial_x u_P$ in terms of $a = \partial_x u_E$ and $b = \partial_x v_E$:

$$s = \frac{a}{a-b}, \tag{2.9}$$

$$0 > \partial_x u_P = a_P = \frac{ab(1+k)}{b+ka}. \tag{2.10}$$

From (2.10) we see that the elastic state (a,b) lies on a hyperbola determined by a_P:

$$(a - c_1)(b - c_2) = c_3, \tag{2.11}$$

where

$$c_1 = \frac{a_P}{1+k}, \quad c_2 = \frac{a_P k}{1+k} = kc_1, \quad c_3 = \frac{a_P^2 k}{(1+k)^2} = c_1 c_2.$$

To classify and understand the elastic-plastic boundaries, we use the following observations repeatedly, relying on (2.7) to express time derivatives in terms of spatial derivatives.

(i) For plastic deformations, $\partial_t u_P > 0$, which implies

$$a_P < 0, \tag{2.12}$$

as in (2.10).

(ii) In an *elastic-to-plastic boundary* $x = st$, the deformation is elastic before the boundary has passed (i.e., for smaller t), and plastic after the boundary has passed (i.e., for larger t). Consequently, in order that the material achieve plastic yield $u = v$ at the boundary, it must be loading elastically: $\partial_t u_E > 0$, which implies $a < 0$.

(iii) In a *plastic-to-elastic boundary* $x = st$, the deformation is plastic before the boundary has passed (i.e., for smaller t), and elastic after the boundary has passed (i.e., for larger t). Consequently, the material must be unloading elastically: $\partial_t u_E < 0$, which implies $a > 0$.

There are significant differences in the description of elastic-plastic boundaries, and in the solution of the SI problem, between the cases $k > 0$ and $k < 0$. *For the remainder of this section, we restrict our attention to the case* $k > 0$. In Appendix A, we return to the case $k < 0$.

From (2.8), (2.12) we draw the following conclusions:

1. $\partial_t u_E = a > 0$ implies $s < (1+k)^{-1}$ or $s > 1$;
2. $\partial_t u_E = a < 0$ implies $(1+k)^{-1} < s < 1$.

We use Fig. 2.1 to distinguish four types of elastic-plastic boundaries, shown in Fig. 2.2. Figure 2.1 shows the (a, b) plane when $k > 0$, with the shaded areas corresponding to constraints on elastic states in each type of elastic-plastic boundary, as explained below. The hyperbola of (2.11) for a fixed value of a_P intersects each of the four shaded regions of Fig. 2.1, as shown.

Type 1. $s > 1$. Fast elastic-to-plastic boundary. The speed and the plastic state are determined by the elastic state (a, b) through relations (2.9), (2.10). From (i), (ii), and $s > 1$ we deduce that $a < b < 0$, shown as region 1 in Fig. 2.2. There is exactly one hyperbola (2.11) through (a, b), and the plastic state $(a_P, a_P), a_P < 0$ is determined by the intersection of the hyperbola with the line $a = b, a < 0$.

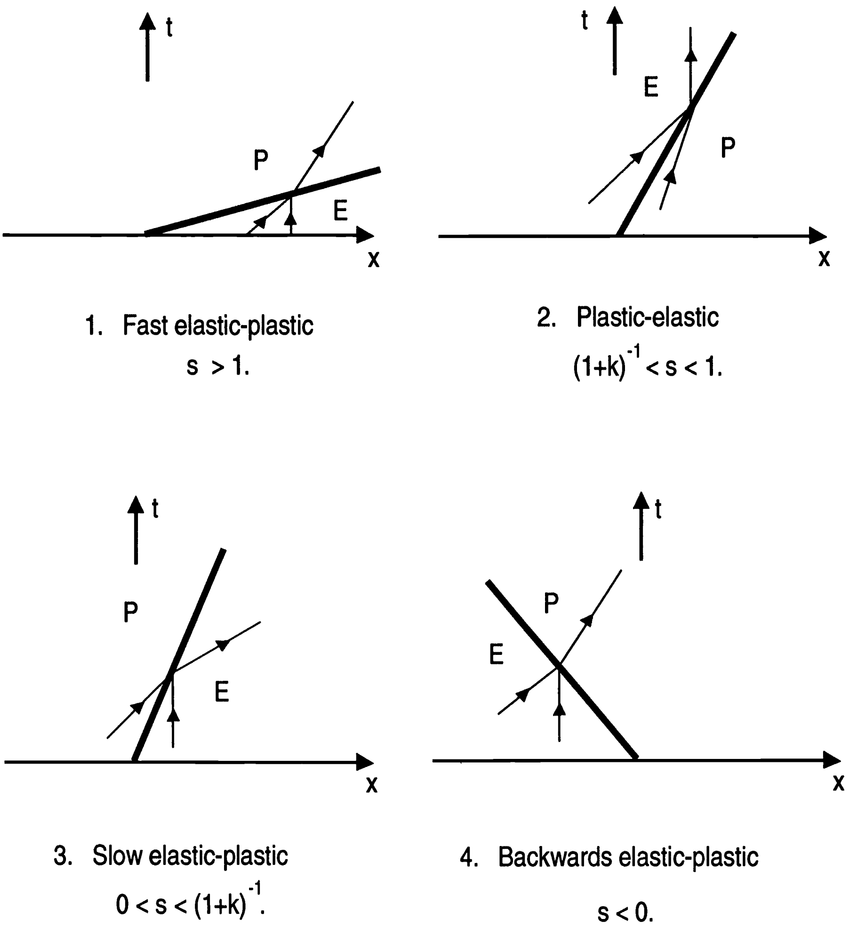
Type 2. $(1+k)^{-1} < s < 1$. Plastic-to-elastic boundary. Here, the plastic state determines a curve of elastic states (a, b). But now the material unloads elastically, so $a > 0$, from (iii) above. The inequalities on s and a_P further imply that $-ka < b < 0$. The corresponding region is labelled 2 in Fig. 2.2. Note that the possible elastic states (a, b) and the corresponding plastic state lie on different arms of the same hyperbola (2.11).

Type 3. $0 < s < (1+k)^{-1}$. Slow elastic-to-plastic boundary. Elastic loading and inequalities on s lead to $0 < -ka < b$ and $a < 0$, shown as region 3 in Fig. 2.2.

Type 4. $s < 0$. Backwards elastic-to-plastic boundary. Elastic loading and inequalities on s lead to $b < a < 0$, shown as region 4 in Fig. 2.2.

In each of the pictures of Fig. 2.1 we have shown the characteristics when $k > 0$. Note that for each boundary exactly two characteristics enter the boundary and one characteristic leaves the boundary. In Fig. 2.3, we show graphs of u, v adjacent to an elastic-plastic boundaries of each of the four types. Note that elastic-plastic boundaries of types 2 and 3 carry a maximum of u and a minimum of v, respectively, whereas u and v are monotonically decreasing across elastic-plastic boundaries of types 1 and 4.

2.2. Solution of the SI problem for $k > 0$. The solution of the SI problem is represented in three ways in Figs. 2.5–2.8.



1. Fast elastic-plastic $s > 1$.

2. Plastic-elastic $(1+k)^{-1} < s < 1$.

3. Slow elastic-plastic $0 < s < (1+k)^{-1}$.

4. Backwards elastic-plastic $s < 0$.

FIG. 2.1. *Elastic-plastic boundaries $x = st$.*

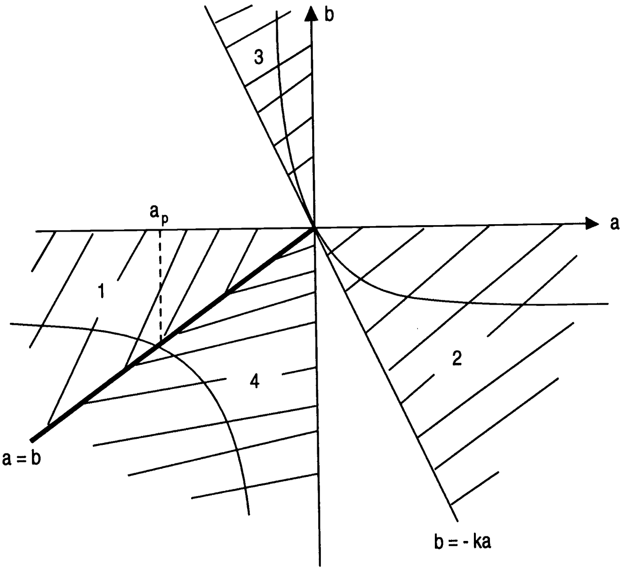


FIG. 2.2. *Location of elastic state (a, b) in elastic-plastic boundaries.*

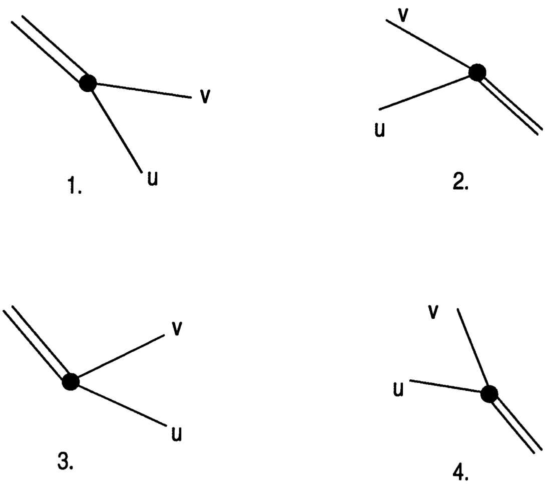


FIG. 2.3. *Graphs of u, v across elastic-plastic boundaries.*

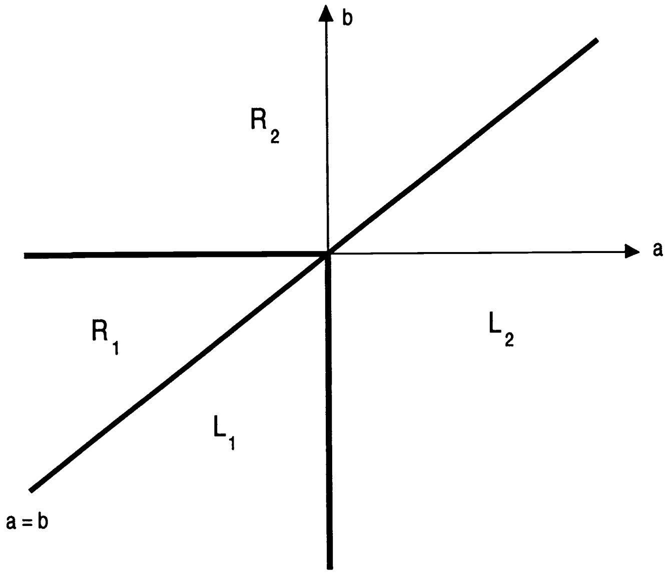


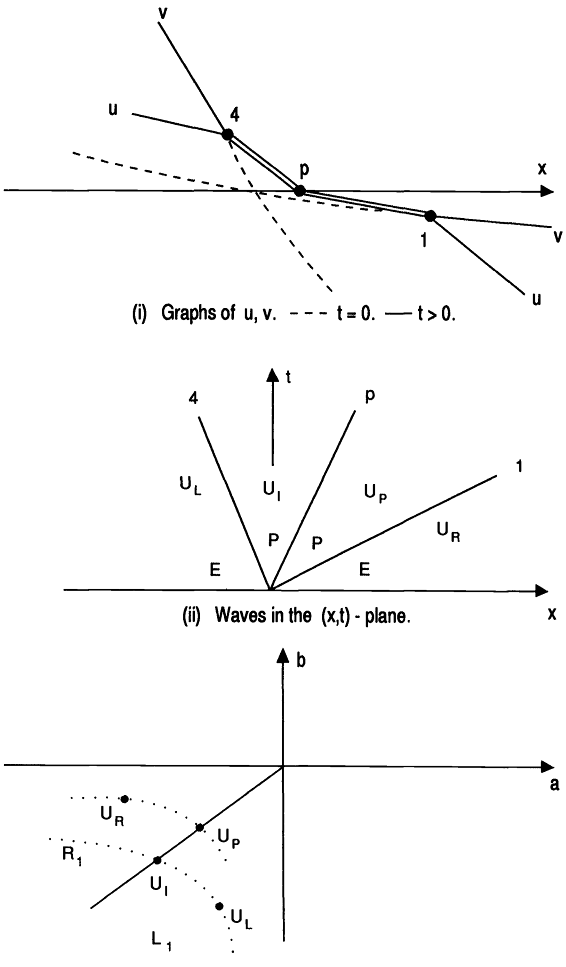
FIG. 2.4. *Location of left and right states.*

(i) Graphs of u and v as functions of x for fixed $t > 0$. Corners in these graphs represent single waves, labelled according to the conventions established above, namely, e for elastic waves, p for plastic waves, and c for contact waves, and a number 1 to 4 to label the type of elastic-plastic boundaries.

(ii) A picture of the (x, t)-plane to show the location of waves and plastic and elastic regions.

(iii) A picture of the (a, b)-plane representing the sequence of values of $\partial_x u$ and $\partial_x v$ between waves. This picture is augmented in the discussion below by formulae for the sequence of points (a, b) between waves.

Since the solution of the SI problem is piecewise linear and scale invariant, it consists of a sequence of single waves joining wedges $\{(x, t) : \alpha < x/t < \beta, t > 0\}$ in



(i) Graphs of u, v. - - - t = 0. —— t > 0.

(ii) Waves in the (x,t) - plane.

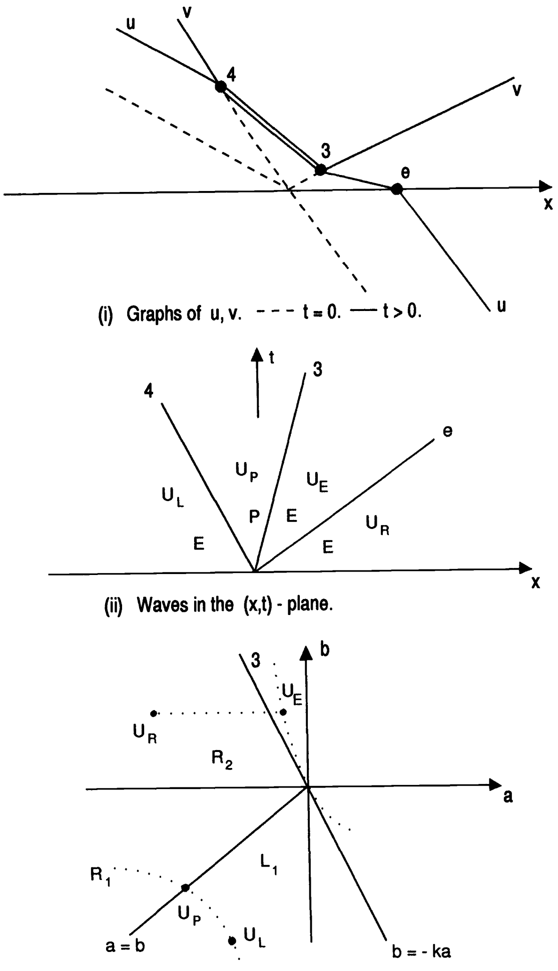
(iii) States in the (a,b) - plane.

FIG. 2.5. *Solution of the* SI *problem. Case* $L_1 R_1$.

the (x,t)-plane. In each wedge the derivatives $\partial_x u, \partial_x v, \partial_t u, \partial_t v$ are all constant. In each wedge we can use the differential equation (1.3) to find the time derivatives in terms of the spatial derivatives, and across each wave, jumps in the time derivatives are determined by jumps in the spatial derivatives, as we have seen. Therefore, the solution may be expressed as a sequence of points in the (a,b)-plane (representing spatial derivatives $(\partial_x u, \partial_x v)$), together with an explanation of the type of single wave separating adjacent wedges in the (x,t)-plane. We refer to each point (a,b) in the sequence as a *state*, since it corresponds to a uniform state (2.3) or (2.5), and we use the notation $U = (a,b)$.

The initial data (1.6) correspond to two points (a_L, b_L), (a_R, b_R) that are the first and last points in the sequence. The types of waves in the sequence vary with the initial data. Because of the restrictions (1.7) on the initial data, $U_L = (a_L, b_L), U_R = (a_R, b_R)$ lie in nonoverlapping regions in the (a,b)-plane, specifically,

$$
\begin{aligned}
a \geq b \quad & \text{for left states} \quad U_L = (a_L, b_L), \\
a \leq b \quad & \text{for right states} \quad U_R = (a_R, b_R).
\end{aligned}
$$



(i) Graphs of u, v. - - - t = 0. —— t > 0.

(ii) Waves in the (x,t) - plane.

(iii) States in the (a,b) - plane.

FIG. 2.6. *Solution of the* SI *problem. Case* L_1R_2.

Each of these regions is further divided into two sectors (see Fig. 2.4).

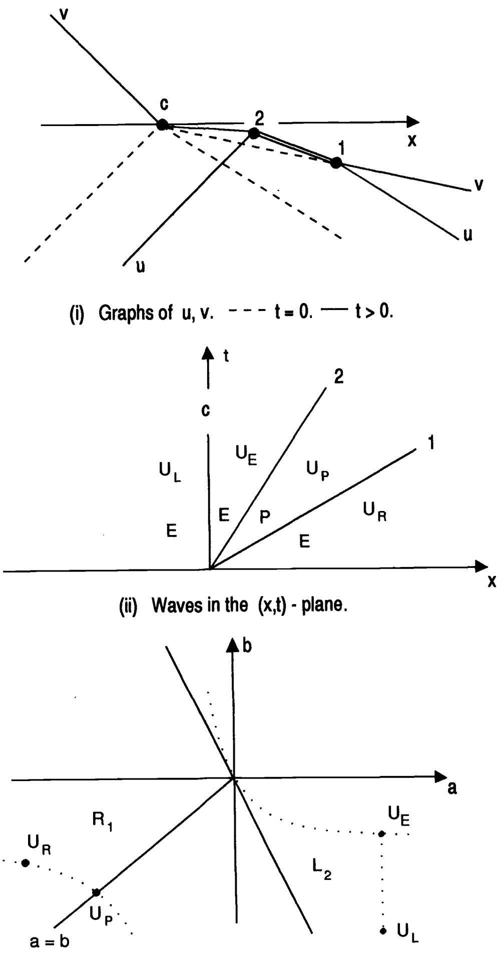
L_1	$b \leq a \leq 0$,
L_2	$b \leq a$ and $a \geq 0$,
R_1	$a \leq b \leq 0$,
R_2	$a \leq b$ and $a \leq 0$.

Given the partition of left and right regions into two sectors, there are four possible combinations of left and right states in different sectors. For each $i = 1$ or $2, j = 1$ or 2, the solution of the SI problem involves the same sequence of single waves for every $U_L \in L_i, U_R \in R_j$.

Case L_1R_1: $U_L = (a_L, b_L) \in L_1$, $U_R = (a_R, b_R) \in R_1$.

Since $(a_R, b_R) \in R_1$, the right state is loading elastically, and reaches plastic yield $u = v$ at a type 1 elastic-plastic boundary

$$x = s_1 t, \qquad s_1 = \frac{a_R}{a_R - b_R} > 1. \tag{2.13}$$



(i) Graphs of u, v. - - - t = 0. —— t > 0.

(ii) Waves in the (x,t) - plane.

(iii) States in the (a,b) - plane.

FIG. 2.7. *Solution of the* SI *problem. Case* L_2R_1.

The plastic state $U_P = (a_P, b_P)$ to the left of this boundary in the (x,t)-plane is also determined by the right state, through (2.10):

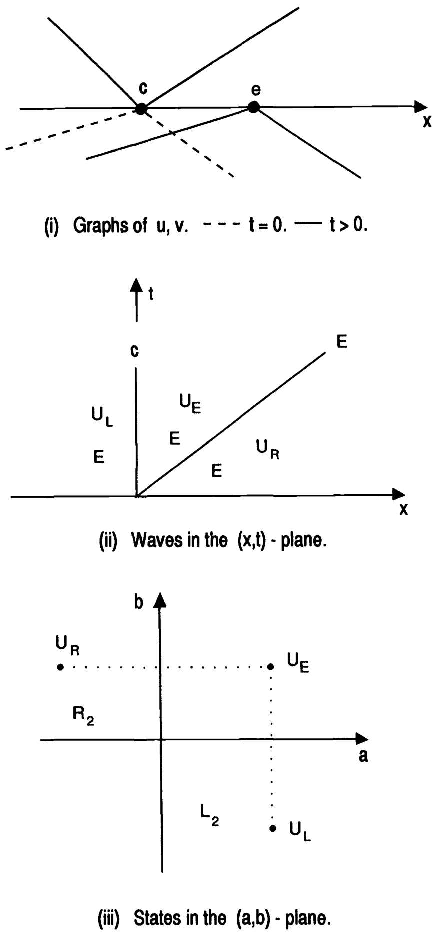
$$a_P = b_P = \frac{a_R b_R (1+k)}{b_R + k a_R}. \tag{2.14}$$

Similarly, since $(a_L, b_L) \in L_1$, the left state loads elastically until it reaches plastic yield at a type 4 elastic-plastic boundary

$$x = s_4 t, \qquad s_4 = \frac{a_L}{a_L - b_L} < 0. \tag{2.15}$$

The plastic state $U_I = (a_I, b_I)$ to the right of this boundary in the (x,t)-plane is also determined by the left state, through (2.10):

$$a_I = b_I = \frac{a_L b_L (1+k)}{b_L + k a_L}. \tag{2.16}$$



(i) Graphs of u, v. - - - t = 0. —— t > 0.

(ii) Waves in the (x,t) - plane.

(iii) States in the (a,b) - plane.

FIG. 2.8. *Solution of the* SI *problem. Case* L_2R_2.

Finally, there is a plastic wave $x = (1+k)^{-1}t$ across which $a = b$ jumps from a_I to a_P.

The solution for $t > 0$ and the corresponding pictures of the (x,t)-plane and (a,b)-plane are shown in Fig. 2.5 for representative choices of $U_L = (a_L, b_L) \in L_1$, $U_R = (a_R, b_R) \in R_1$. In Fig. 2.5(i) we show the solution itself with corners (i.e., waves) marked by dots, labelled according to the classification of single waves. The material is deforming plastically where the two graphs coincide, indicated by double lines. Note that the slopes of the different line segments, while all negative, need not be in the relations suggested by the figure. For example, the slope at P may jump down or up, depending on the location of U_L and U_R. This is most easily seen in Fig. 2.5(iii): clearly in this figure, U_L and U_R may be chosen in L_1 and R_1, respectively, so that U_I lies closer to the origin than U_P, in which case the slope at P would jump down. Letting $\xi = x/t \to c_P = (1+k)^{-1}$ in (2.2), we see that the plastic corner labelled p necessarily lies on the x axis.

Case L_1R_2: $U_L = (a_L, b_L) \in L_1$, $U_R = (a_R, b_R) \in R_2$.

Since $(a_L, b_L) \in L_1$, the left state loads elastically until it reaches plastic yield at a type 4 elastic-plastic boundary given by (2.15), with the plastic state $U_I = (a_I, b_I)$ on the right given by (2.16).

The right state $U_R = (a_R, b_R)$ may or may not be loading elastically (depending on the sign of a_R), but in any case, it remains elastic and uniform up until $x = t$, i.e., in the wedge $0 < t \leq x$. The ray $x = t$ is then an elastic wave, and U_R can jump to a new state $U_E = (a_E, b_R)$. (Recall that b does not change across an elastic wave.)

We now have to reconcile a determined plastic state U_P to the left in the (x,t)-plane of an elastic state U_E depending on a free parameter a_E. The only available wave is a plastic-to-elastic boundary of type 2. Indeed, U_E must lie in the region labelled 3 in Fig. 2.4, and be such that U_E lies on the hyperbola of (2.11), with $a_P = a_I$:

$$\frac{ab(1+k)}{b+ka} = a_I. \tag{2.17}$$

Setting $b = b_R$, and solving (2.17) for $a = a_E$, with a_I given by (2.16), we get

$$a_E = \left(k\left(\frac{1}{b_L} - \frac{1}{b_R}\right) + \frac{1}{a_L}\right)^{-1}. \tag{2.18}$$

Case L_2R_1: $U_L = (a_L, b_L) \in L_2$, $U_R = (a_R, b_R) \in R_1$.

As in case L_1R_1, the right state is loading elastically, and reaches plastic yield $u = v$ at a type 1 elastic-plastic boundary given by (2.13), with the corresponding plastic state $U_P = (a_P, a_P)$ given by (2.14).

The left state has $a_L > 0$. Therefore, the left state is unloading elastically and remains uniform for all $t > 0$. The line $t = 0$ is a contact wave across which $b = \partial_x v$ jumps. The elastic state immediately to the right of this wave in the (x,t)-plane is $U_E = (a_L, b_E)$. The parameter b_E is determined by compatibility with the plastic state U_P. The only available elastic-plastic boundary is of type 2. From equation (2.10), with $(a,b) = U_E = (a_L, b_E)$, and equation (2.14), we obtain

$$b_E = k\left(\frac{k}{b_R} + \frac{1}{a_R} - \frac{1}{a_L}\right)^{-1}. \tag{2.19}$$

Case L_2R_2: $U_L = (a_L, b_L) \in L_2$, $U_R = (a_R, b_R) \in R_2$.

Here, the solution is purely elastic. There is a contact wave on the t-axis and an elastic wave $x = t$. The intermediate elastic state is $U_E = (a_L, b_R)$.

Finally, it is easy to check that the solution is continuous with respect to the data. In particular, if U_L or U_R lies on the boundary of one of the regions, then the solution can be obtained as a limit of solutions with initial data from any of the adjacent regions.

3. Solution of the Cauchy problem.

In this section we prove existence and uniqueness results for the pure initial value problem (Cauchy problem):

$$\begin{gathered} \partial_t u + k\partial_t v + \partial_x u = 0, \quad -\infty < x < \infty, \quad t > 0, \\ \partial_t v = \begin{cases} \partial_t u & \text{if} \quad u = v \quad \text{and} \quad \partial_t u \geq 0, \\ 0 & \text{otherwise,} \end{cases} \end{gathered} \tag{3.1}$$

with continuous initial data (u_o, v_o):

$$(u,v)(x,0) = (u_o, v_o)(x), \qquad -\infty < x < \infty. \tag{3.2}$$

A continuous function $(u,v)(x,t)$ is a *solution* (more precisely, a *weak solution*) of the Cauchy problem (3.1), (3.2) if

$$\iint_{\mathbb{R}^2} \{(u(x,t)+kv(x,t))\phi_t(x,t)+u(x,t)\phi_x\}\,dx\,dt=0 \tag{3.3}$$

for each C^1 function $\phi(x,t)$ with compact support in the upper half plane $t>0$, if

$$v(x,t)=\max\left\{v_o(x),\ \max_{0\le s\le t} u(x,s)\right\},\quad -\infty<x<\infty,\quad t>0, \tag{3.4}$$

and if

$$(u,v)(x,0)=(u_o,v_o)(x),\qquad -\infty<x<\infty. \tag{3.5}$$

We remark that if a piecewise C^1 continuous function (u,v) satisfies (3.1) almost everywhere, and satisfies (3.2), then it is a solution.

The proofs of existence and uniqueness depend upon a careful analysis of solutions of initial value problems in which the initial data are piecewise linear and continuous. In particular, we study the propagation and interaction of discontinuities in the derivatives $\partial_x u, \partial_x v$ of the solution. Much of this section is devoted to studying piecewise linear continuous solutions of (3.1).

3.1. Piecewise linear solutions. For our purposes a function $f:\mathbb{R}^{>}\to\mathbb{R}^{K}$ will be called *piecewise linear* on a subset Ω of $\mathbb{R}^{>}$ if it is continuous on Ω, and there are a finite number of disjoint open sets $\Omega_j\subset\Omega, j=1,\dots,k$, with $\Omega=\overline{\cup_{j=1}^k\Omega_j}$ such that f is affine on each Ω_j. (Note that each Ω_j has polygonal boundary.) Our main goal in this subsection (§3.1) is to prove the following existence result for piecewise linear initial data. This is the main step in proving the general existence theorem for general continuous bounded initial data (see Theorem 3.11).

THEOREM 3.1. *If* $u_o(x)\le v_o(x)$ *are piecewise linear and continuous, and are constant outside a bounded interval, then there is a global solution of the initial value problem* (3.1), (3.2) *that is piecewise linear and continuous in the upper half plane.*

The proof of this result depends on an analysis of functionals that describe the propagation of discontinuities in derivatives of the solution. We begin by introducing some terminology in §3.1.1. In §3.1.2, we define and analyze the functionals, in a process we call *corner counting.* In §3.1.3, we describe the detailed structure of piecewise linear solutions of system (3.1), and in §3.1.4, we complete the proof of Theorem 3.1. Section 3.1.5 is concerned with properties of the solution operator, which takes given piecewise linear initial data to the solution given in Theorem 3.1. These properties are used in §3.2, in which it is shown that the piecewise linear solutions converge to a continuous solution of (3.1), (3.2).

3.1.1. Terminology. Let $U=(u(x,t),v(x,t))$ be a solution of (3.1) for $(x,t)\in\mathbb{R}\times[0,T_o]$ for some $T_o>0$. Assume that U is piecewise linear and continuous on $\mathbb{R}\times[0,T_o]$. Consider a fixed value t_o of t in $[0,T_o]$. An open interval $J=(x_1,x_2)$ (which may be unbounded) is called an *elastic interval* if $(u,v)(x,t_o)$ represents an elastic deformation for all $x\in J$. That is, J is an elastic interval if $u(x,t_o)<v(x,t_o)$ or $\partial_t u(x,t_o)<0$ for all x in J. Similarly, $J=(x_1,x_2)$ is called a *plastic interval* if $(u,v)(x,t_o)$ represents a plastic deformation for all $x\in J$: i.e., $u(x,t_o)=v(x,t_o)$ and $\partial_t u(x,t_o)\ge 0$ for all x in J. (Equality is allowed in the latter inequality to include the degenerate case of neutral loading, in which u and v are equal to the same constant

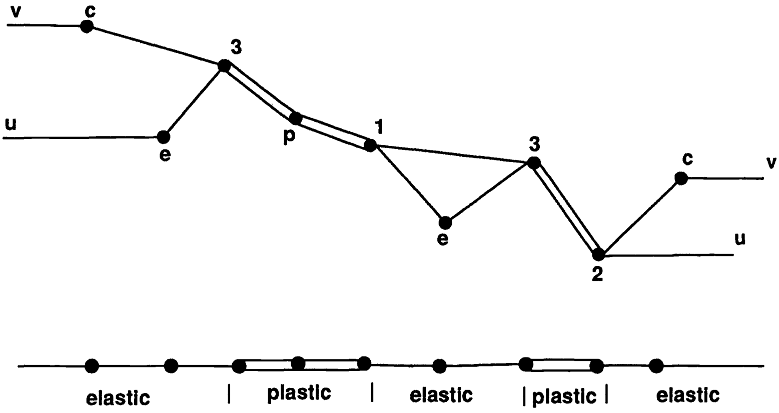


FIG. 3.1(a). *Graphs of piecewise linear u, v illustrating regular points, elastic intervals and plastic intervals.* $F(t) = 9$.

in the interval J.) We shall always assume that elastic intervals and plastic intervals are *maximal* in the sense that they are not contained in larger intervals of the same type.

The solution (u, v) has a *corner* at $x = x_o$ (we also say x_o is a corner) if either $\partial_x u(x, t_o)$ or $\partial_x v(x, t_o)$ is discontinuous at $x = x_o$. A corner x_o is called an *elastic-plastic boundary* if x_o is an endpoint of an elastic interval. As in §2, we shall often refer to a curve in the (x, t)-plane as an elastic-plastic boundary if it is the union of points (x_o, t_o), where x_o is an elastic-plastic boundary at time t_o. A characteristic curve $x = \gamma(t)$ is called a *corner characteristic* if there is a corner at $x = \gamma(t)$ for each t. We shall often speak of corner characteristics at a fixed time, meaning the corner at that time, which lies on a corner characteristic. A corner characteristic is called *monotone* if both $u(x, t)$ and $v(x, t)$ are monotone in x across the corner for each t. If the corner is a local extremum for u or v, then the corner characteristic is called *nonmonotone*.

In Fig. 3.1(a), we show examples of graphs of $u(x, t_o), v(x, t_o)$ at fixed t_o, with regular corners of various types defined in §2. We also show the elastic intervals and plastic intervals. The only nonmonotone corner characteristics in these examples involve a minimum of u at one of the corner characteristics labelled e.

Since $(u(x, t), v(x, t))$ is assumed to be piecewise linear, there are a finite number of corners at each time $t \in [0, T_o]$, and each corner lies on a corner characteristic, or is an elastic-plastic boundary, or is the origin of a local SI problem. (Here we use the term *local* to indicate that the solution is scale invariant only locally in space and time.) A point (x_o, t_o), with $t_o < T_o$, is called a *regular corner point* if there is exactly one corner characteristic or elastic-plastic boundary containing (x_o, t_o). If there is a corner at $x = x_o, t = t_o$ that is not a regular corner point, then it is the origin of a nontrivial local SI problem; such a point will be called an *intersection point.* We shall occasionally refer to intersection points as *collision points* (because they involve the collision of two or more corners or of the graph of u with the graph of v) or as *interaction points* (because, thinking of corners as acceleration waves, intersection points give rise to wave interactions).

3.1.2. Corner counting. We proceed to construct a global solution of the initial value problem for piecewise linear continuous and bounded initial data $(u_o(x), v_o(x))$

satisfying $u_o(x) \leq v_o(x)$ for all x.

Initially, there may be some finite number of SI problems to solve. Then over some time interval, there are no interactions. That is, the solution has corner characteristics and elastic-plastic boundaries, but these curves do not intersect, and moreover, $u < v$ in each elastic interval. Let $T > 0$ be the maximum time for which there are no interactions for $0 < t < T$.

At time T, there are interactions, which form centers for SI problems. The solution proceeds by solving these SI problems.

The proof of Theorem 3.1 depends upon an analysis of functionals that at each time measure the number of corners, corner characteristics, and extrema in the solution. We define these functionals as follows. Our definitions are guided by the functionals used by Glimm in the context of the Cauchy problem for hyperbolic systems of conservation laws [2].

Let $n(t)$ be the number of corners at time t, and let $L(t)$ denote the number of corner characteristics. If both $\partial_x u$ and $\partial_x v$ jump at a corner characteristic in an elastic interval, then count each separately (i.e., such a corner characteristic contributes 2 to $L(t)$).

Finally, we count the number of *approaching pairs.* First define an *approaching pair* at time t to be a pair of corners x_o, x_1 such that $x_o < x_1$ and u have a local maximum at x_o while v has a local minimum at x_1. Let $Q(t)$ be the number of approaching pairs at time t. Define

$$F(t) = L(t) + 2Q(t). \tag{3.6}$$

We illustrate these quantities and their relationship to each other and the structure of solutions by appealing to Fig. 3.1. In Fig. 3.1(a), we have $n(t) = 9, L(t) = 5, Q(t) = 2, F(t) = 9$. (The points labelled c, e, p each contribute 1 to $L(t)$; each point labelled 3 contributes 1 to $Q(t)$; each of the nine corners contributes 1 to $n(t)$.) In Fig. 3.1(b), we have taken the same sequence of elastic and plastic intervals, but have perturbed the graphs of u and v to show a solution with fewer corners. Specifically, in Fig. 3.1(a), one of the corners in the left-hand semi-infinite elastic interval labelled c or e may be removed (in Fig. 3.1(b), we retained only e), and the plastic corner p may be removed. Then $n(t) = 7, L(t) = 3, Q(t) = 2, F(t) = 7$. In an extreme case, we could have neutral loading (i.e., $u = v$) in both semi-infinite elastic intervals. This would reduce the number of corner characteristics by two, giving $n(t) = 5, L(t) = 1, Q(t) = 2, F(t) = 5$. We illustrate the extreme case in Fig. 3.1(c).

This discussion suggests that we can estimate the number of corners $n(t)$ in the solution by $3L(t) + 2$, which involves only the number of corner characteristics. But this latter number generally increases with time. However, we may obtain a nonincreasing upper bound for this estimate by including the term $Q(t)$ that counts the number of approaching pairs. This term is analogous to the term measuring *the potential for future interaction* in Glimm's analysis of conservation laws [2]. In fact, we show in Theorem 3.2 that $n(t)$ is bounded by $3F(t) + 2$, and then go on to show in Lemma 3.5 that $F(t)$ is nonincreasing in time t. We then have a global upper bound ((3.8) below) for the number of corners in the solution. This is a major step in showing that the solution may be continued, and remains piecewise linear, for all time.

The estimate $3L(t) + 2$ for $n(t)$ is justified loosely as follows. As suggested by our discussion of Fig. 3.1, each elastic interval contains a corner characteristic, providing we adopt the convention that neutral loading is plastic. (This result is proved

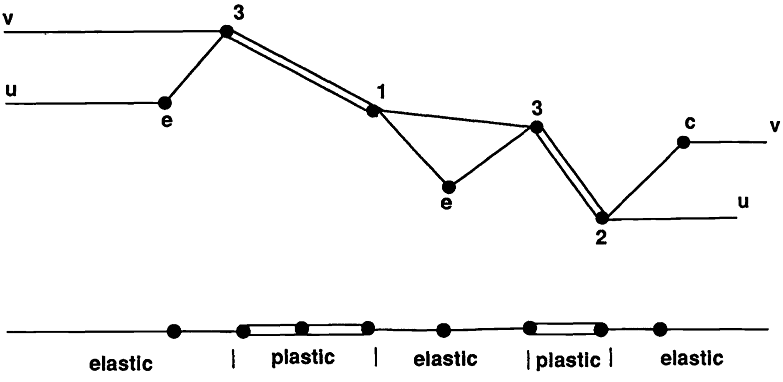


FIG. 3.1(b). *Graphs of piecewise linear u, v illustrating regular points, elastic intervals and plastic intervals.* $F(t) = 7$.

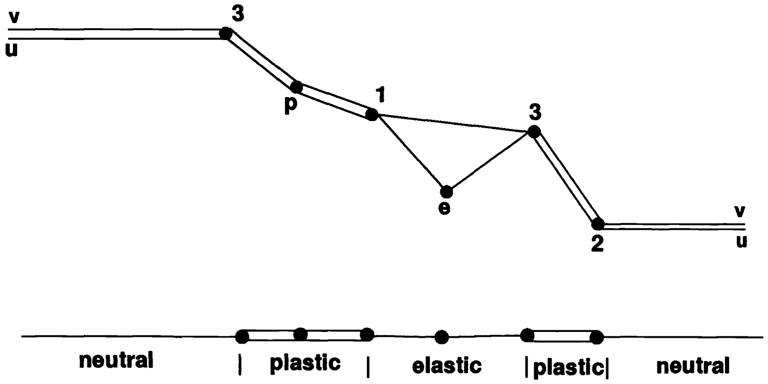


FIG. 3.1(c). *Graphs of piecewise linear u, v illustrating regular points, elastic intervals and plastic intervals.* $F(t) = 5$.

as Lemma 3.3 below.) Moreover, any corner that is not a corner characteristic is an elastic-plastic boundary. Therefore, we get an upper bound for the number of corners, in terms of the number of corner characteristics, by assuming that all corner characteristics (whether elastic or plastic or contact) are embedded in a bounded elastic interval containing no other corner characteristics. The endpoints of this interval, which are elastic-plastic boundaries, contribute another two corners. In this way, each corner characteristic counts as three in our estimate of the number of corners in the solution. Now, if there are corner characteristics in the semi-infinite intervals (as in Fig. 3.1(b)), then we overestimate the number of elastic-plastic boundaries in this counting, and hence $n(t) < 3L(t)$, but if both semi-infinite intervals have neutral loading, i.e., with no corners, and are adjacent to plastic intervals, then we have missed two elastic-plastic boundaries. Therefore, we must add two to our estimate of the number of corners in the solution. In particular, the configuration in Fig. 3.1(c) is minimal in the sense that this estimate is achieved: $n(t) = 3L(t) + 2$. We formalize the estimate in Theorem 3.2 below.

THEOREM 3.2. *Let $(u(x,t), v(x,t))$ be a piecewise linear continuous solution of* (1.3) *for $-\infty < x < \infty$, t in an open interval I. Then*

$$\begin{aligned} n(t) &\le 3L(t) + 2 \\ &\le 3F(t) + 2. \end{aligned} \tag{3.7}$$

The proof depends upon the following observation, which we state as a lemma.

LEMMA 3.3. *Let $(u(x,t), v(x,t))$ be a piecewise linear continuous solution of* (1.3) *for $-\infty < x < \infty$, t in an open interval I. At a fixed $t \in I$, every bounded elastic interval contains a corner characteristic. Moreover, if $\partial_x u$ and $\partial_x v$ have bounded support, then any semi-infinite elastic interval also contains a corner characteristic.*

Proof. We argue by contradiction. Let $J = (x_1, x_2)$ be a bounded elastic interval at time t_o that does not contain a corner characteristic. Then $u(x,t_o)$ and $v(x,t_o)$ are affine for $x \in J$. But the endpoints of J are elastic-plastic boundaries, so that $u(x_1,t) = v(x_1,t)$ and $u(x_2,t) = v(x_2,t)$. Consequently, $u(x,t) = v(x,t)$ for all $x \in J$. Since J is an elastic interval, we have $\partial_x u = -\partial_t u > 0$, and $\partial_x u$ is constant. Moreover, $\partial_t v = 0$ in J. It follows that there is a subinterval K of J, and $\tau < t$ in I such that

$$u(x,\tau) > u(x,t) = v(x,t) = v(x,\tau) \quad \text{for all } x \in K.$$

But this is impossible, since $u(x,t) \leq v(x,t)$ for all (x,t).

A similar argument applies to semi-infinite elastic intervals when $\partial_x u$ and $\partial_x v$ have bounded support. Specifically, if J is such an interval, then the finite endpoint is an elastic-plastic boundary, so $u = v$ there. Since u and v cannot be equal and constant throughout an elastic interval, and they have to be constant outside some finite interval, there must be a corner characteristic somewhere in J. □

Proof of Theorem 3.2. The number of corners in the closure of a bounded elastic interval is less than or equal to the number of corner characteristics, plus two, as discussed above with reference to Fig. 3.1. It then follows from Lemma 3.3 that this number is less than or equal to three times the number of corner characteristics. The only corners in a plastic interval are the corner characteristics, so that the number of corners in each such interval is also less than or equal to three times the number of corner characteristics. Finally, we add 2 to this formula to admit the possibility that two unbounded elastic intervals have elastic-plastic boundary points that have not yet been counted as corners. This proves the first inequality in (3.6). The second inequality is trivial, since $Q(t) \geq 0$. □

The proof of Theorem 3.1 depends upon using solutions of SI problems to analyze the behavior of the functionals $L(t), Q(t), F(t)$.

LEMMA 3.4. *$Q(t)$ is a nonincreasing function of time.*

Proof. As a first step in the proof, we note that no critical points in $u(x,t)$ or $v(x,t)$ (as functions of x for fixed t) are created through collisions. This is because each intersection point (x_o,t_o) is the center of a local SI problem, for which the solution for $t > t_o$ (given in §3) has at most one critical point in u and v. Therefore, the solution for $t < t_o$ near (x_o,t_o) has at least the same number of critical points in the same variables. (For example, if $u(.,t)$ has a maximum for $t > t_o$, then $\partial_x u > 0$ for x to the left of the maximum, and $\partial_x u < 0$ for x to the right of the maximum for $t > t_o$. But then $\partial_x u > 0$ for small x, and $\partial_x u < 0$ for large x for $t < t_o$, so that $u(.,t)$ also has at least one maximum for $t < t_o$.)

Now maxima of u are carried by elastic corner characteristics or type 2 elastic-plastic boundaries; therefore, maxima of u travel with speed *at least* $(1+k)^{-1}$, the plastic wave speed. On the other hand, minima of v are carried by contact corner characteristics or type 3 elastic-plastic boundaries; therefore, minima of v travel with speed *at most* $(1+k)^{-1}$. Consequently, once a maximum of u is to the right of a minimum of v, it remains so for all future time. Since no critical points are created, $Q(t)$ cannot increase. □

The functionals $L(t), Q(t), F(t)$ change only due to intersections. Therefore, we can understand the global behavior of $F(t)$ by analyzing how $L(t), Q(t)$ change locally near an intersection point. We shall speak of changes of the functionals *across* an intersection point (x_o, t_o), meaning the effect of that intersection point on changes in the functionals as t crosses t_o.

More precisely, this local analysis can be performed by assuming that the only corners in the solution are carried by curves entering or emanating from the intersection point. That is, when discussing an intersection point in isolation, we shall consider $L(t), Q(t)$ to be defined by considering the solution only in a small neighborhood of that point, i.e., $L(t)$ is the number of corner characteristics in the neighborhood at time t, and $Q(t)$ is the number of approaching pairs in the neighborhood at time t. The neighborhood of an intersection point (x_o, t_o) can clearly be defined so that $L(t)$ and $Q(t)$ are constant for $t < t_o$, and for $t > t_o$.

LEMMA 3.5. *$F(t)$ is nonincreasing.*

Proof. We need only show that $F(t)$ is nonincreasing across intersection points. Let (x_o, t_o) be an intersection point. Then it is the center of an SI problem. Therefore, the data and solution fall into one of the four cases $L_iR_j, \quad i, j = 1, 2$ of §2. In all cases we have $Q(t) = 0$ after the intersection. Except for the case L_2R_2, we have $L(t) \leq 1$, and hence $F(t) \leq 1$, after the intersection. Apart from this case, we have $L(t) \geq 1$ before intersection, since if $L(t) = 0$, then either there is no intersection, or two elastic-plastic boundaries pinch off a plastic region, which is necessarily case L_2R_2. (Specifically, an elastic-plastic boundary of type 2, carrying a maximum of u converges from the left on an elastic-plastic boundary of type 3, carrying a minimum of v. All other intersections of elastic-plastic boundaries pinch off elastic intervals, which necessarily contain a corner characteristic.) Since $Q(t) \geq 0$, we see that $F(t) \geq 1$ before intersection and since $F(t) \leq 1$ after the intersection, we have completed the proof except in the case L_2R_2.

In case L_2R_2, we have $L(t) \leq 2, Q(t) = 0$ after the intersection, so that $F(t) \leq 2$. Therefore, if $Q(t) \geq 1$ before the intersection, we have $F \geq 2$ before the intersection and the proof is complete. Therefore, suppose $Q(t) = 0$ before the intersection. But in case L_2R_2 we have $\partial_x u_L > 0$ and $\partial_x v_R > 0$. Thus, before the intersection, either u has a maximum and v has a minimum, or at least one of u and v is monotonic increasing. In the former case, since $Q(t) = 0$, the maximum of u must be to the right of the minimum of v. But then it stays on the right, and there can be no intersection of these corners. For the other possibility, namely, one of $u(x,t), v(x,t)$ being monotonic increasing in x, this means that the deformation is elastic. If $\partial_x u > 0$, then from (1.3), $\partial_t u < 0$, so $u(x,t)$ is decreasing in time, and the deformation remains elastic, with no intersection. If $\partial_x v > 0$, then since the graph of $u(x,t)$ is to the right of the graph of $v(x,t)$, it remains to the right, the deformation remains elastic, and again there is no intersection. We have shown that $Q(t) \geq 1$ in Case L_2R_2 before the intersection point. This completes the proof. □

The consequence of Theorem 3.2 and Lemma 3.5 is that the number of corners $n(t)$ is bounded as long as the solution can be constructed by solving SI problems and using propagation along characteristics:

$$n(t) \leq 3F(t) + 2 \leq 3F(0) + 2. \tag{3.8}$$

Another immediate consequence of Lemma 3.5 is the following property.

LEMMA 3.6. *If $L(t)$ increases across an intersection point, then $2Q(t)$ decreases across that intersection point by at least the same amount.*

Proof. The proof of Lemma 3.5 demonstrated that $F(t)$ decreases across intersection points. The result now follows immediately from the definition of $F(t)$. □

The solution of the initial value problem with piecewise linear continuous initial data continues by solving SI problems at intersection points. This process can continue as long as there are a finite number of intersection points. Let T^* be the maximal time with the property that the solution is defined, and has a finite number of intersection points for $-\infty < x < \infty, 0 \leq t \leq T$, for each $T, 0 < T < T^*$. We now consider the solution defined on the region $-\infty < x < \infty, 0 \leq t < T^*$. Within this region, intersection points are still isolated in space-time, but of course they may in principle accumulate at some point on the boundary, with $t = T^*$. (Because u and v are constant outside an interval, the intersection points cannot approach $x = \pm\infty$ in finite time.)

LEMMA 3.7. *There are a finite number of intersection points across which L changes. If L and Q are constant across an intersection point, then $L(t) = 1$ and $Q(t) = 0$ across the intersection point.*

Proof. Lemma 3.4 implies that there are a finite number of intersection points at which Q decreases, since $Q(t)$ is a nonincreasing nonnegative integer valued function. Therefore, by Lemma 3.6, there are a finite number of points at which $L(t)$ increases. Since $L(t)$ is a nonnegative integer valued function, there can only be a finite number of intersection points across which $L(t)$ can decrease. Therefore, there are a finite number of intersection points across which L changes.

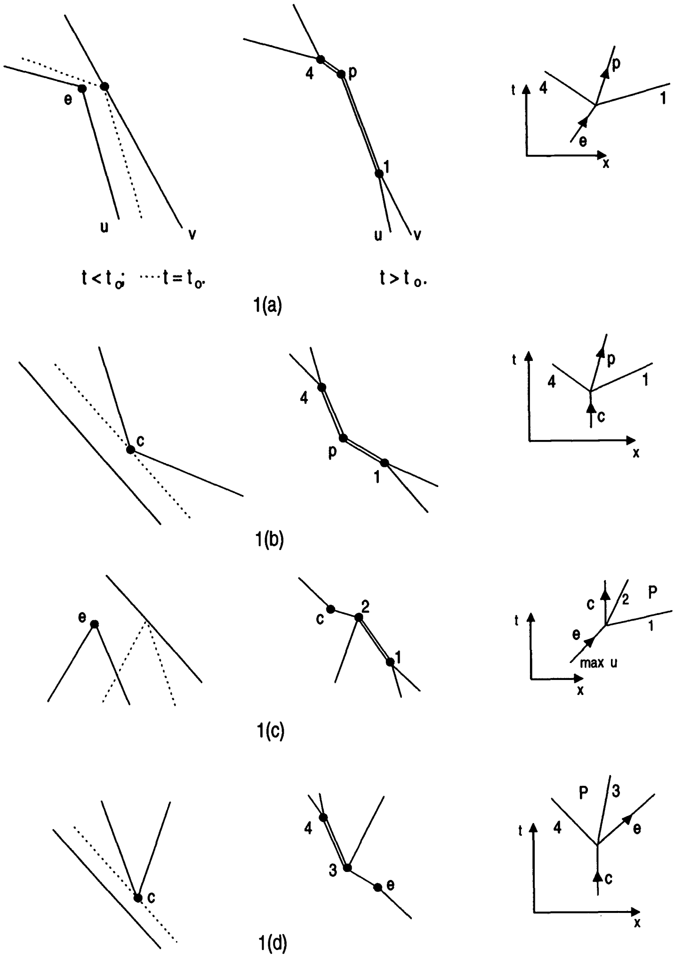
To show the second statement of the lemma, we argue as in the proof of Lemma 3.5. Considering intersection points as centers of SI problems, note that case L_2R_2 is ruled out because $Q(t)$ necessarily decreases across the intersection point in that case. Therefore, we have $L(t) \leq 1$ and $Q(t) = 0$ after the intersection. Lemma 3.4 then implies $Q(t) = 0$ across the intersection point. If $L(t) = 0$ is constant across the intersection point, then either there is no intersection or we are again in case L_2R_2. Therefore, $L(t) = 1$ across the intersection point. □

We have shown that after a certain time $T < T^*$, every intersection point has $L(t) = 1$ and $Q(t) = 0$ across the intersection point. We refer to such intersection points as *generic*. There is a finite list of possible generic intersection points for $T < t < T^*$, as we now discuss.

3.1.3. Generic collisions.

Consider intersection points having the property that $L(t) = 1, Q(t) = 0$ are constant across the intersection point. Because $L(t) = 1$, there is exactly one corner characteristic entering the intersection point, and exactly one corner characteristic leaving the intersection point.

Suppose the entering corner characteristic is plastic. Then at each time it lies in a plastic interval. If this plastic interval is bounded, then the endpoints both approach the intersection point, and form elastic-plastic boundaries that pinch off the plastic region containing the corner characteristic. But the only pair of approaching elastic-plastic boundaries pinching off a plastic region involves a type 2 front to the left of a type 3 front. In this case, we have seen that $Q(t) = 1$ and is, therefore, ruled out by the requirement $Q(t) = 0$. On the other hand, if there were any bounded elastic intervals, whose endpoints approach the intersection point, then by Theorem 3.2 there would be additional elastic corner characteristics, giving $L(t) \geq 2$. We conclude that if a plastic corner characteristic enters the intersection point, then there is at most one elastic-plastic boundary entering the intersection point. Since the plastic corner characteristic and the elastic-plastic boundary approach, we must have a type 2 front to the left of the corner characteristic, or a type 3 front to the right of the corner



1(a)

1(b)

1(c)

1(d)

FIG. 3.2 (1). *First touch.*

characteristic. These are the only possible intersections involving a plastic corner characteristic for which $L(t) = 1$ and $Q(t) = 0$.

If the entering corner characteristic is elastic or a contact, then a similar argument leads to the conclusion that there can be at most two elastic-plastic boundaries entering the intersection point. For otherwise, a pair of fronts would pinch off a plastic region, which was just ruled out because this implies $Q(t) \geq 1$.

Using these observations, we compile an exhaustive list of intersection points having the property that $L(t) = 1$ and $Q(t) = 0$ across the intersection point. This list is represented in Fig. 3.2. In this figure, we also include the additional case in which $Q(t)$ decreases and $L(t)$ increases keeping $F(t)$ constant.

Remarks. We conjecture that Fig. 3.2 includes all the possible collisions of corners in piecewise linear continuous solutions that cannot be altered qualitatively by small perturbations of the solution, that is, that generic intersection points are generic in the usual sense. We also conjecture that Fig. 3.2 includes all the possible intersection points in piecewise linear continuous solutions across which $F(t)$ is constant.

We divide the collisions into four categories as follows.

1. First touch. An elastic corner characteristic reaches a point of plastic yield ($u =$

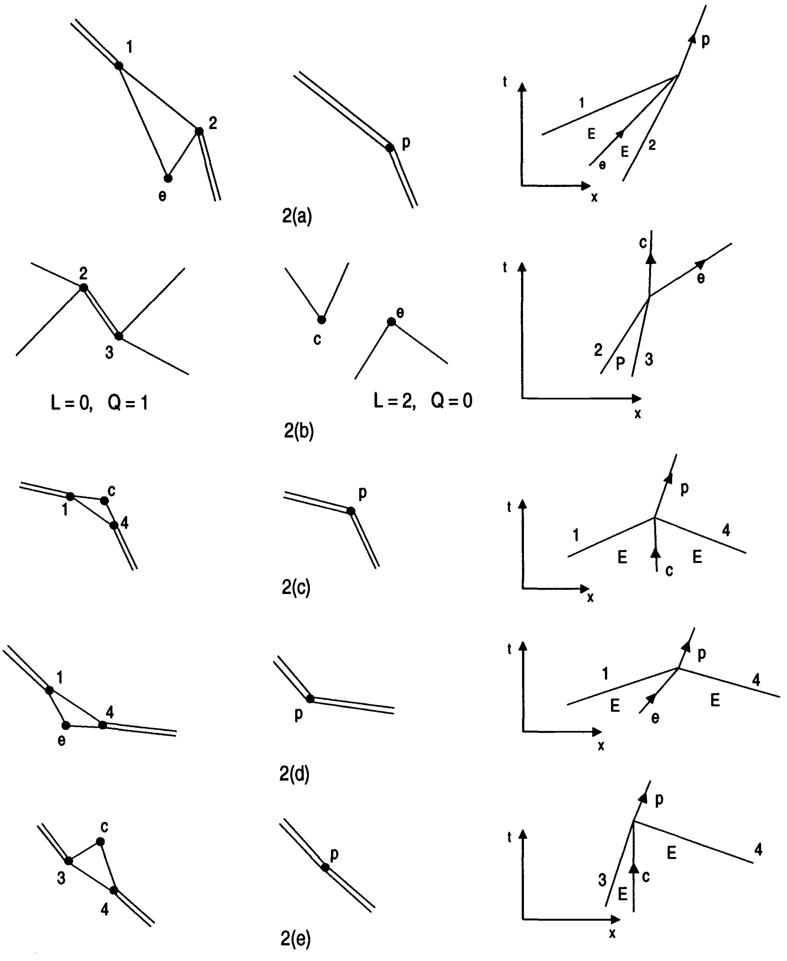


FIG. 3.2 (2). *Collisions of elastic-plastic boundaries.*

v) at the corner, and one of the variables is smooth (i.e., linear) at the intersection.

2. Collision of two elastic-plastic boundaries. Some such collisions involve corner characteristics.

3. Collision of an elastic-plastic boundary and a monotone corner characteristic.

4. Collision of an elastic-plastic boundary and a nonmonotone corner characteristic.

In Fig. 3.2 we show graphs of u, v before, at, and after collision, and the lines in the (x, t)-plane representing elastic-plastic boundaries and corner characteristics. This list of "generic" collisions is the basis for much of the analysis of this section, so we spend some time interpreting the information of Fig. 3.2.

Here are some properties that we read from Fig. 3.2, referring only to the intersection points shown there.

1. $L(t)$ and $Q(t)$ are unchanged across every collision except one (Fig. 3.2.2b), in which $L(t)$ increases by 2 while $Q(t)$ decreases by 1. $F(t)$ is unchanged in each case.

2. In case 2, the collision of elastic-plastic boundaries, an elastic region is pinched off except in case 2(b), where a plastic region is pinched off.

3. In case 3, in which an elastic-plastic boundary collides with a monotone corner

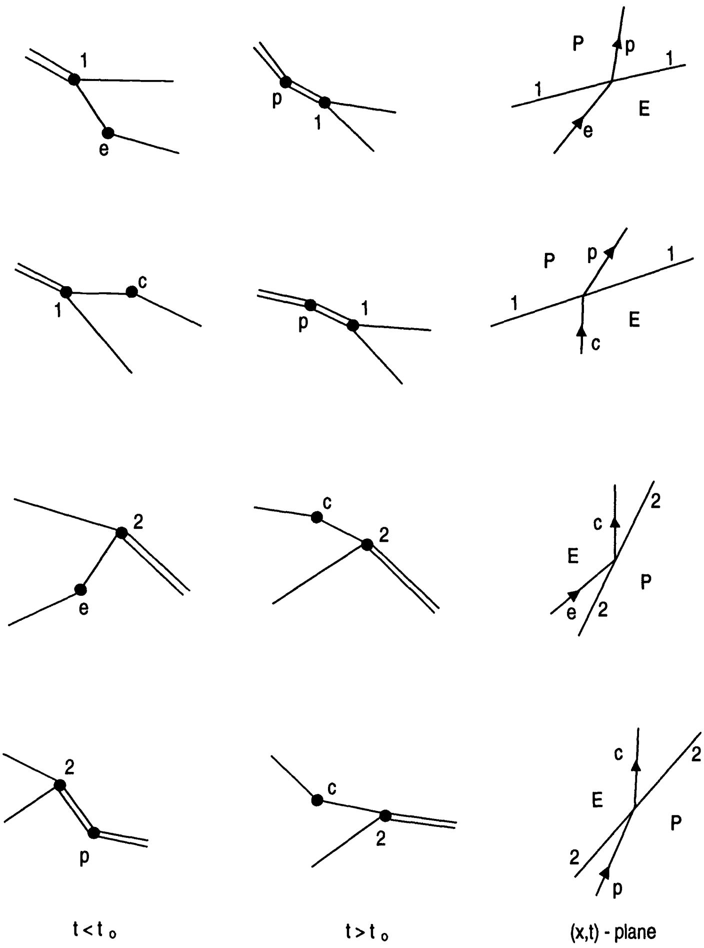


FIG. 3.2 (3). (*Types* 1, 2.) *Elastic-plastic boundary, monotone corner characteristic.*

characteristic, the type of the boundary is unchanged, and the corner characteristic remains monotone.

4. In case 4, in which an elastic-plastic boundary collides with a nonmonotone corner characteristic, the type of the boundary is changed, and the corner characteristic becomes monotone.

5. Only monotonic corner characteristics leave intersection points. In particular, if a nonmonotone corner characteristic enters an intersection point, then a monotone corner characteristic leaves.

6. All intersection points involving monotone corner characteristics and monotone elastic-plastic boundaries (of type 1 or 4) have $\partial_x u < 0$ and $\partial_x v < 0$ in a neighborhood of the intersection point.

3.1.4. Proof of Theorem 3.1. Consider the structure of a solution that is piecewise linear up to some maximal time T^*. By (3.8), the number of corners, and hence the number of corner characteristics, is bounded above at each time t in the interval $[0, T^*)$. Since T^* is maximal, there must be an infinite number of intersection points in $(-\infty, \infty) \times [0, T^*)$; for otherwise the solution could be continued, and remain piecewise linear, past T^*. But Lemma 3.7 states that after a finite number

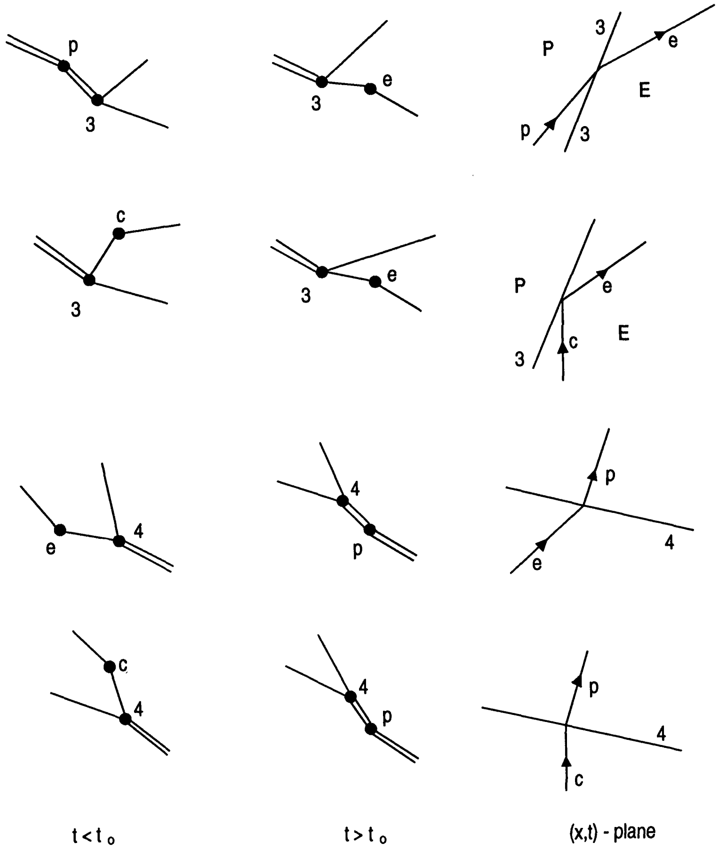


FIG. 3.2 (3). (*cont.*).(*Types* 3, 4.) *Elastic-plastic boundary, monotone corner characteristic.*

of intersections, i.e., after some time $T < T^*$, all intersection points have $L(t) = 1, Q(t) = 0$. Thereafter, corner characteristics are neither created nor destroyed, and their interactions occur only according to the list in Fig. 3.2.

Let us trace a single corner characteristic C as it propagates forward in time $t > T$ through intersection points listed in Fig. 3.2. By property 5, if C is nonmonotonic, either there are no intersection points on C, or it becomes monotonic after the first intersection point. In discussing further intersection points on C, we are, therefore, restricted to those in Fig. 3.2 involving only monotone corner characteristics. Regarding these, there are only two cases in which a plastic corner characteristic enters an intersection point.

(a) An elastic-plastic boundary of type 2 also enters the intersection point, in which case the corner characteristic emerges as a contact corner characteristic with

$$\partial_x u > 0 \quad \text{and} \quad \partial_x v < 0.$$

(b) An elastic-plastic boundary of type 3 also enters the intersection point, in which case the corner characteristic emerges as an elastic corner characteristic with

$$\partial_x u < 0 \quad \text{and} \quad \partial_x v > 0.$$

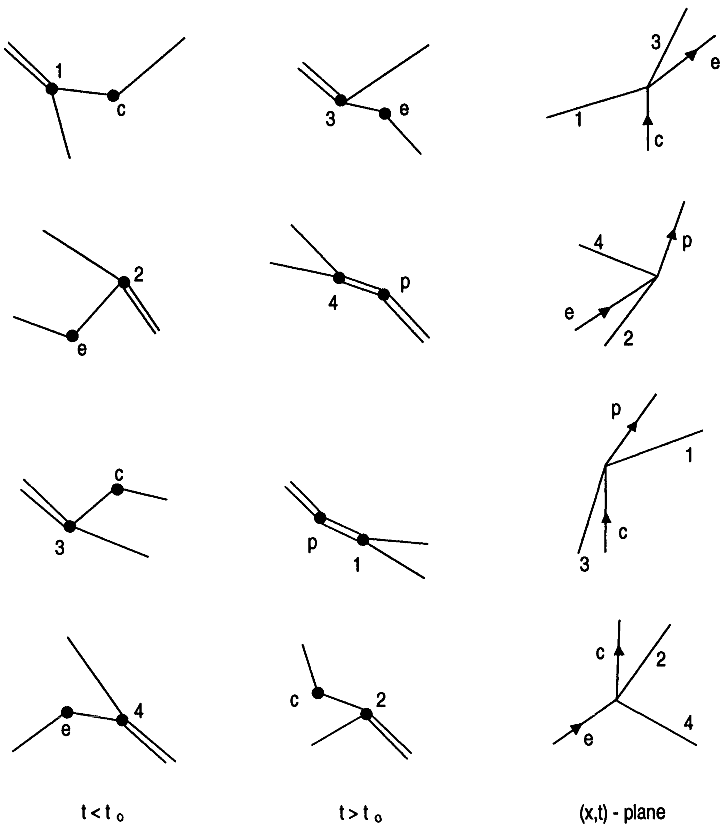


FIG. 3.2 (4). *Elastic-plastic boundary, nonmonotone corner characteristic.*

Moreover, the remaining intersections involving type 2 or type 3 elastic-plastic boundaries may be summarized as follows.

(c) An elastic corner characteristic enters a type 2 elastic-plastic boundary from the left, and emerges as a contact corner characteristic, also on the left. In the elastic region, in particular along the corner characteristics, we have

$$\partial_x u > 0 \quad \text{and} \quad \partial_x v < 0.$$

(d) A contact corner characteristic enters a type 3 elastic-plastic boundary from the right, and emerges as an elastic corner characteristic, also on the right. In the elastic region, in particular along the corner characteristics, we have

$$\partial_x u < 0 \quad \text{and} \quad \partial_x v > 0.$$

Suppose a portion of C is a plastic corner characteristic. Then the next intersection point on C must be one of (a) or (b); therefore, C becomes either a contact corner characteristic with $\partial_x u > 0$ and $\partial_x v < 0$, or an elastic corner characteristic with $\partial_x u < 0$ and $\partial_x v > 0$. In either case, there can be no further intersections along C, because neither of (c) or (d) is consistent with these possibilities, and intersection with a monotone elastic-plastic boundary of type 1 or 4 is ruled out by property 6 above.

Suppose on the other hand that C never becomes a plastic corner characteristic. The only intersections in which neither the incoming nor the outgoing monotone corner characteristics are plastic are given by (c) and (d). But then the outgoing characteristics are inconsistent with the possibility of further intersection points.

We conclude that there are at most two intersection points on C for $t > T$. Therefore, there are at most a finite number of intersection points in the solution for $t < T^*$, so they cannot accumulate. We have shown that $T^* = \infty$, and the solution remains piecewise linear in (x, t) for all time. This concludes the proof of Theorem 3.1. □

3.1.5. Properties of the solution operator. We first show that piecewise linear continuous solutions of the Cauchy problem with piecewise linear continuous initial data depend monotonically on the initial data in the following sense. We put the obvious partial order on pairs of functions: $U_o(x) = (u_o(x), v_o(x)) \leq U_1(x) = (u_1(x), v_1(x))$ if $u_o(x) \leq u_1(x)$ and $v_o(x) \leq v_1(x)$ for all x. Let $S(t)$ denote the solution operator:

$$S(t)U_o(x) = (u(x,t), v(x,t)).$$

We say that $S(t)$ is *monotonically increasing* if $U_o \leq U_1$ implies $S(t)U_o \leq S(t)U_1$.

THEOREM 3.8. *$S(t)$ is monotonically increasing on piecewise linear continuous functions.*

Proof. We begin by showing that solutions of the scale invariant problem depend monotonically on the initial data. For the SI problem, note that increasing $(u_o(x), v_o(x))$ is the same as decreasing a_L and b_L, and/or increasing a_R and b_R.

Case L_2R_2 is trivial. In the other cases, we note that corner characteristics are pinned on the x-axis at the appropriate value of x determined by the characteristic speed. For a plastic corner characteristic, both u and v are zero at this point, while for an elastic (respectively, contact) corner characteristic, only u (respectively, v) is zero at this point. This property follows by letting $\xi = x/t$ approach the characteristic speed in (2.3) or (2.5).

The strategy of the proof is to first increase a_R, b_R, and show that the solution functions u, v both increase, and then to decrease a_L, b_L, showing again that the solution functions u, v both increase.

In case L_1R_1 (see Fig. 2.5), the elastic-plastic boundary of type 1 is independent of the elastic-plastic boundary of type 4, because the plastic corner characteristic in the solution is fixed. Increasing a_R, b_R clearly increases a_P, while decreasing a_L, b_L clearly decreases a_I. Both of these correspond to an increase in u, v.

Consider cases L_1R_2 (see Fig. 2.6) and L_2R_1 (see Fig. 2.7). First note that a_E is decreased in case L_1R_2, and b_E is increased in case L_2R_1 by increasing a_R, b_R, or decreasing a_L, b_L. But a_P is decreased in case L_1R_2, and increased in case L_2R_1. These observations lead to the conclusion that u and v increase in both cases when the initial data are increased.

By continuity, the same conclusion holds when the data are degenerate, corresponding to the boundary between two of the L_iR_j cases. Now consider the action of $S(t)$ on general piecewise linear continuous functions. Any points of intersection of graphs of u or of v for different data may be treated as SI problems, for which we have established monotonicity. It now follows that $S(t)$ is monotonically increasing on piecewise linear continuous functions. We omit the details of the argument. □

Next we show that $S(t)$ is a contraction with respect to the sup norm on piecewise linear continuous functions. We define the supremum norm (or L_∞ norm) on piecewise linear continuous functions $U(x) = (u,v)(x)$ as follows:

$$||U|| = \sup_x \max\{u(x), v(x)\}.$$

THEOREM 3.9. *$S(t)$ is an L_∞ contraction on piecewise linear continuous functions.*

Proof. The proof relies on the following lemma, reflecting an obvious property of solutions.

LEMMA 3.10. *Let $(u_1, v_1), (u_2, v_2)$ be two piecewise linear continuous solutions of the Cauchy problem. If the solutions differ by a constant α at time $t = 0$,*

$$u_1(x,0) - u_2(x,0) = \alpha, \qquad v_1(x,0) - v_2(x,0) = \alpha,$$

then they differ by the same constant for all $t \geq 0$:

$$u_1(x,t) - u_2(x,t) = \alpha, \qquad v_1(x,t) - v_2(x,t) = \alpha.$$

Proof of Theorem 3.9. Consider two sets of initial data $U_o^i(x)$, $i = 1, 2$, and let $U^i(x,t) = S(t)U_o^i(x)$ be the corresponding solutions of the Cauchy problem.

Let $\alpha = ||U_o^2 - U_o^1||$. Then by monotonicity we have the following inequalities for $t > 0$, since they are true for $t = 0$.

$$\begin{aligned} u_1 - \alpha \leq u_2 \leq u_1 + \alpha, \\ v_1 - \alpha \leq v_2 \leq v_1 + \alpha, \end{aligned}$$

i.e., $||U^1(.,t) - U^2(.,t)|| = ||S(t)U_o^1 - S(t)U_o^2| \leq ||U_o^2 - U_o^1||$, as required. □

3.2. Convergence to a continuous solution. We can now prove the following existence result.

THEOREM 3.11. *Let the initial data $(u_o(x), v_o(x))$ be bounded and continuous, and satisfy $u_o(x) \leq v_o(x)$ for each x. Then there is a solution $(u,v)(x,t)$ of the Cauchy problem* (1.3) *that is bounded and continuous.*

Proof. Let (u_o^n, v_o^n) denote the continuous approximation of the initial data (u_o, v_o), defined as follows:

$$(u_o^n, v_o^n)\left(\frac{j}{n}\right) = (u_o, v_o)\left(\frac{j}{n}\right), \qquad -n^2 \leq j \leq n^2.$$

(u_o^n, v_o^n) is linear on each subinterval $j/n \leq x \leq (j+1)/n$, and constant outside the interval $(-n^2, n^2)$. Then

$$\begin{aligned} & u_o^n(x) \leq v_o^n(x), \qquad -\infty < x < \infty \\ & \sup_x |u_o^n(x)| \leq \sup_x |u_o(x)|, \\ & \sup_x |v_o^n(x)| \leq \sup_x |v_o(x)|. \end{aligned} \tag{3.9}$$

Moreover, it is easy to see that $\{(u_o^n, v_o^n)\}$ converges uniformly to (u_o, v_o) on every interval $[-N, N]$.

Let $(u^n, v^n)(x,t) = S(t)(u_o^n, v_o^n)(x)$ denote the solution of the Cauchy problem with initial data $(u_o^n, v_o^n)(x)$. Since $S(t)$ is a contraction, we have

$$||S(t)(u_o^n, v_o^n) - S(t)(u_o^m, v_o^m)|| \leq ||(u_o^n, v_o^n) - (u_o^m, v_o^m)||.$$

But $\{(u^n_o, v^n_o)\}$ is uniformly convergent, so $(u^n, v^n)(x,t)$ satisfies the uniform Cauchy criterion as a function of both x and t, and hence converges uniformly in t as well as in x to a uniformly continuous function $(u,v)(x,t)$.

To see that (u,v) is a solution of the Cauchy problem, we simply take limits in the weak formulation. That is, each member of the sequence satisfies (3.2)–(3.4), and by uniform convergence, we can take the appropriate limits, and conclude that (u,v) satisfies equations (3.2)–(3.4) also. □

3.3. Uniqueness of entropy solutions. Let $X = BUC(\mathbb{R})$ be the space of bounded uniformly continuous functions with norm $||f||_X = \sup_x |f(x)|$, and let $Y = X \times X$ with norm $||(u,v)|| = ||u||_X + ||v||_X$. Define $Z = BUC([0,\infty), Y)$, with norm $||U||_Z = \sup_0 ||U(t)||_Y$. An *entropy solution* of the Cauchy problem (3.1) is a solution that can be approximated in Z by piecewise linear continuous solutions over finite time intervals. Let $S(t)$ denote the unique continuous extension of $S(t)$ to the closed subset $\{(u,v) \in Y : u(x) \le v(x), -\infty < x < \infty\}$ of Y.

THEOREM 3.12. *Let the initial data $(u_o(x), v_o(x))$ be in the space Y, and satisfy $u_o(x) \le v_o(x)$ for each x. Then there is exactly one entropy solution $(u,v)(x,t)$ of the Cauchy problem* (1.3).

Proof. Let U, V be two entropy solutions with the same initial data U_o, and let U^n, V^n be the corresponding piecewise linear approximations of U, V, respectively. Then $U^n_o(x) = U^n(x,0), V^n_o(x) = V^n(x,0)$ both converge uniformly to $U_o(x)$ and $U^n(x,t) = S(t)U^n_o(x)$, $V^n(x,t) = S(t)V^n_o(x)$. Therefore,

$$U(.,t) = \lim_{n\to\infty} S(t)U^n_o = S(t)U_o = \lim_{n\to\infty} S(t)V^n_o = V(.,t).$$

This completes the proof. □

A. Appendix. Negative *k*.

In our treatment [5] of the SI problem for the piecewise linear longitudinal model ((1.1), with k constant), we investigated the implications of negative k in some detail. For the unidirectional model of this paper, we discuss two of the mathematical issues raised in [5]. Specifically, we demonstrate nonexistence and nonuniqueness of solutions of the SI problem (1.3)–(1.6) if $-1 < k < 0$. (Recall that $k > -1$ ensures that plastic and elastic waves propagate with positive speed. For the longitudinal model (1.1), $k > -1$ ensures hyperbolicity.)

A.1. Nonuniqueness for $k < 0$. First note that $-1 < k < 0$ implies that $c_E = 1 < c_P = (1+k)^{-1}$. That is, elastic characteristics are slower than plastic characteristics. This has little effect on elastic-plastic boundaries of types 1,3,4 but has a profound effect on those of type 2. For elastic-plastic boundaries of types 1,3,4, the u-characteristics (composed of elastic and plastic characteristics, see §3.1.3) pass through the front, while the contact characteristics enter. This is true for all $k > -1$. For elastic-plastic boundaries of type 2, however, the contact characteristics leave the front, and the u-characteristics enter from both sides if $k > 0$, but *leave* the front from both sides if $-1 < k < 0$. Thus, for $-1 < k < 0$, elastic-plastic boundaries of type 2 are unstable in the sense of Lax [3]. We call these *entropy-violating fronts.* A consequence of allowing entropy-violating fronts is that there are multiple solutions of some SI problems. In fact we shall construct two-parameter families of solutions of SI problems with the same initial data.

First we show how the construction of elastic-plastic boundaries in §2 must be modified when $-1 < k < 0$. The hyperbola of (2.11) still describes possible elastic states adjacent to an elastic-plastic boundary for a given plastic state. The hyperbola

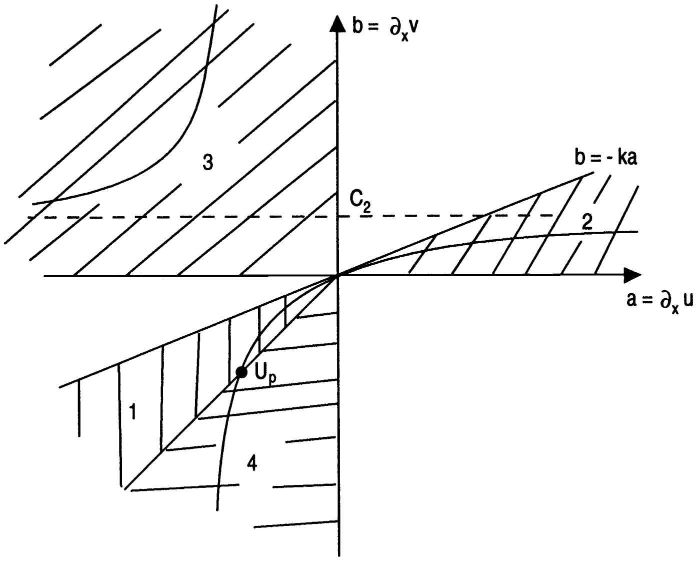


FIG. A.1. $k < 0$.

is redrawn in Fig. A.1 for $-1 < k < 0$, together with the various regions corresponding to elastic-plastic boundaries of types 1–4. (This figure is the analogue of Fig. 2.2, in which $k > 0$.)

Consider a plastic right state $U_R = U_P$, and let U_L lie on the portion of the hyperbola in region 2, as shown in Fig. A.1. Then one solution of the SI problem is an elastic-plastic boundary of type 2, shown in Fig. A.2. Since plastic characteristics move faster than the front, while elastic and contact characteristics move more slowly, it is not surprising that we can insert additional waves ahead of and behind the front, and adjust the speed of the front, without changing the initial data. These solutions are shown in Fig. A.3(a), with the corresponding construction shown in Fig. A.3(b). Here, U_R is joined to an arbitrary plastic state U_I. Then U_I is joined to an arbitrary point U_E in Region 2 on the hyperbola through U_I. The point U_E now determines the elastic and contact characteristic fronts. U_E is joined to the point U_C shown in the figure by an elastic wave, and U_C is joined to U_L by a contact corner characteristic. Each of these solutions involves an entropy-violating front, unless U_I and U_E are at the origin. In that case, the construction and solution are shown in Fig. A.4. Here, there is a neutrally loaded region, in which $\partial_x u = \partial_x v = 0$, and the material is in both the elastic and the plastic state. This is the preferred solution of the SI problem, since it does not contain an entropy-violating front.

A.2. Nonexistence for $k < 0$. Now we turn to the issue of nonexistence of solutions of SI problems for some choices of initial conditions. In Fig. A.5, we have divided the (a, b)-plane into regions, somewhat as we did in Fig. 2.4 for $k > 0$.

We proceed by varying U_R. If U_R is in Region R_1, then there is an elastic-plastic boundary of type 1, like there is for $k > 0$. However, as U_R approaches the ray labelled OA, the speed of this front approaches $c_P = (1 + k)^{-1}$, and U_P goes to infinity. Since the solution (u, v) remains bounded near the front, this suggests that the solution is tending towards a jump discontinuity.

For U_R in Region R_2, there can be no continuous solution of the SI problem. The

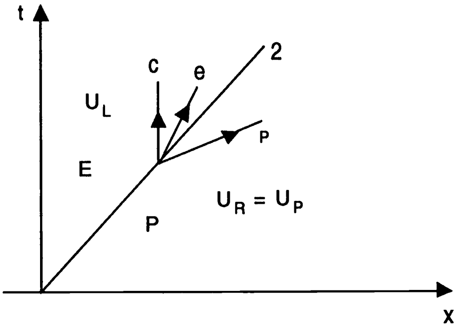


FIG. A.2. *Entropy-violating front.*

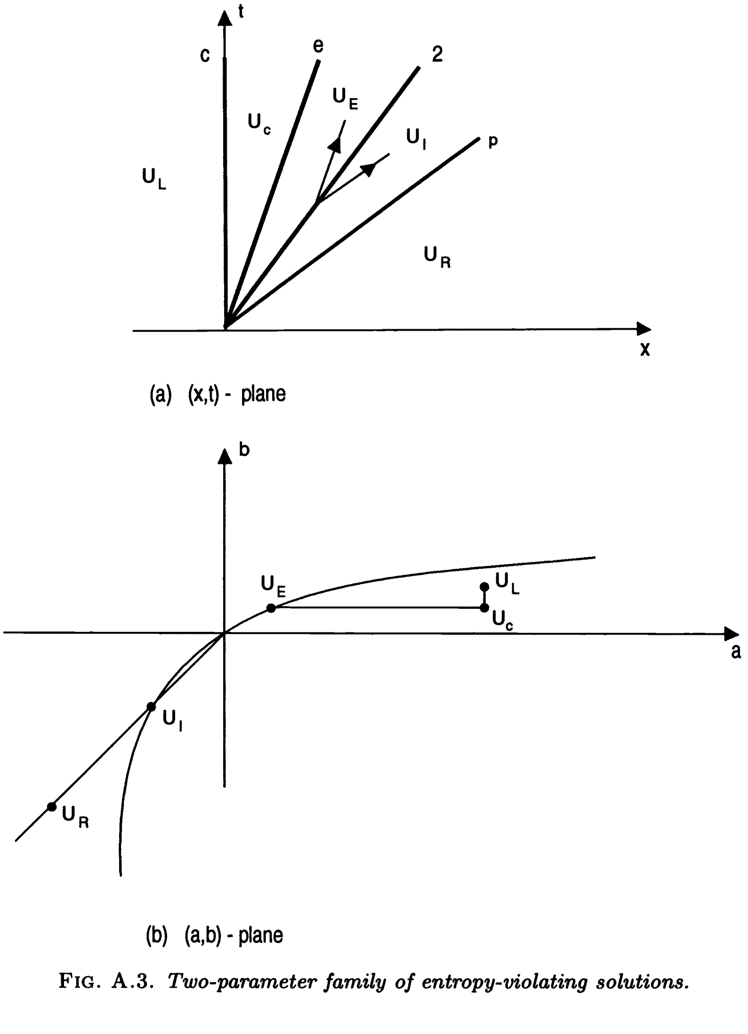


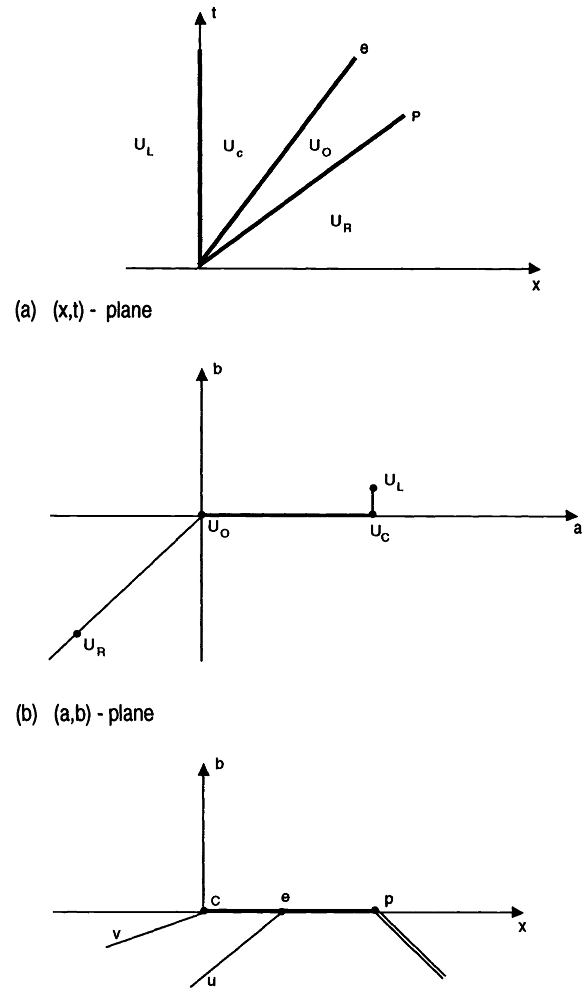
FIG. A.3. *Two-parameter family of entropy-violating solutions.*

elastic state reaches yield along a curve $x = st$, where s is given by (2.9), and satisfies

$$c_E = 1 < s < (1+k)^{-1} = c_P$$

in the wedge R_2:

$$-ka_R < b_R < 0.$$



(a) (x,t) - plane

(b) (a,b) - plane

(c) Solution graphs for t > 0.

FIG. A.4. *Entropy solution.*

However, the plastic state $U_P = (a_P, a_P)$ to which U_R would be connected is given by (2.14):

$$a_P = b_P = \frac{a_R b_R(1+k)}{b_R + ka_R} > 0, \tag{A.1}$$

which violates the condition (2.12) for plastic deformation; therefore, there can be no solution if U_R is in region R_2. However, the behavior of the continuous solution as U_R approaches region R_2 from within region R_1 suggests that there may be discontinuous solutions of the SI problem when U_R is in region R_2. We plan to investigate discontinuous solutions in a future paper.

If U_R is in Region R_3, then there can be a type 3 elastic-plastic boundary, like there is for $k > 0$. However, for U_L in Region L_1, the plastic state U_P is determined; therefore, the hyperbola in the construction of the type 3 elastic-plastic boundary is

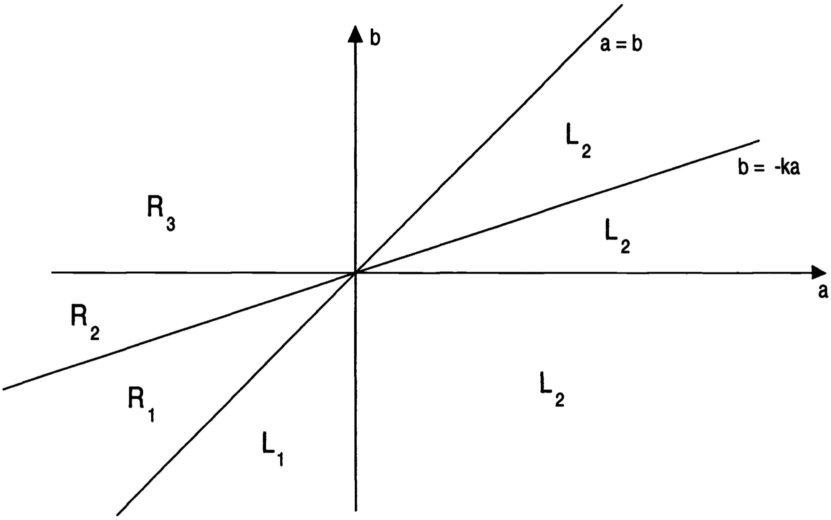


FIG. A.5. *The (a, b)-plane for $k < 0$.*

fixed by U_L. For U_R above the horizontal asymptote of this hyperbola,

$$a_R > c_1 = \frac{a_P k}{a_P + b_P},$$

there is no problem, since a faster elastic characteristic can be used to move $\partial_x U$ onto the hyperbola. But for $a_R < c_1$, there can be no global solution because no matter which elastic front is used, the elastic state does not lie on a portion of the hyperbola defined by U_L corresponding to elastic-to-plastic fronts.

Finally, for U_L in Region L_2, we saw in §A.1 that an apparent problem with nonuniqueness of solutions is resolved by imposing the entropy condition. But there is also an unresolved problem with existence similar to the situation for U_L in Region L_1.

REFERENCES

[1] S. S. ANTMAN AND W. G. SZYMCZAK, *Nonlinear elastoplastic waves*, Contemp. Math., 100 (1989), pp. 27–54.

[2] J. GLIMM, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.

[3] P. D. LAX, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[4] E. H. LEE, *A boundary value problem in the theory of plastic wave propagation*, Quart. Appl. Math., 10 (1953), pp. 335–346.

[5] D. G. SCHAEFFER AND M. SHEARER, *Scale-invariant initial value problems in one dimensional elastoplasticity, with consequences for multidimensional nonassociative plasticity*, European J. Appl. Math., 3 (1992), pp. 225–254.

[6] T. C. T. TING, *Nonexistence of higher order discontinuities across elastic/plastic boundary in elastic-plastic wave propagation*, in Plasticity and Failure Behavior of Solids, G. C. Sih, A. J. Ishlinsky, and S. T. Mileiko., eds., Kluwer Academic Press, Hingham, MA, 1990.

[7] J. A. TRANGENSTEIN AND R. B. PEMBER, *The Riemann problem for longitudinal motion in an elastic-plastic bar*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 180–207.

SIAM J. MATH. ANAL.
Vol. 24, No. 5, pp. 1145–1152, September 1993

002

A LEVEL SET FORMULATION FOR THE SOLUTION OF THE DIRICHLET PROBLEM FOR HAMILTON–JACOBI EQUATIONS*

STANLEY OSHER†

Abstract. A level set formulation for the solution of the Hamilton–Jacobi equation $F(x, y, u, u_x, u_y) = 0$ is presented, where u is prescribed on a set of closed bounded noncharacteristic curves. A time dependent Hamilton–Jacobi equation is derived such that the zero level set at various time t of this solution is precisely the set of points (x, y) for which $u(x, y) = t$. This gives a fast and simple numerical method for generating the viscosity solution to $F = 0$. The level set capturing idea was first introduced by Osher and Sethian [*J. Comput. Phys.*, 79 (1988), pp. 12–49], and the observation that this is useful for an important computer vision problem of this type was then made by Kimmel and Bruckstein in [*Technion* (*Israel*) *Computer Science Report*, CIS #9209, 1992] following Bruckstein [*Comput. Vision Graphics Image Process*, 44 (1988), pp. 139–154]. Finally, it is noted that an extension to many space dimensions is immediate.

Key words. Hamilton–Jacobi equation, viscosity solution, level set, numerical method

AMS subject classifications. 35L99, 65M05

Introduction. We are interested in solving a general first-order partial differential equation for a function $z = u(x, y)$ of the type

$$F(x, y, z, p, q) = 0, \tag{0.1}$$

where $p = u_x$, $q = u_y$.

This is a classical problem in partial differential equations (P.D.E.). In fact, the method of characteristics was invented to solve it. Typically we are given Cauchy data on a curve Γ, i.e., for

$$x = x_0(s), \qquad y = y_0(s); \tag{0.2a}$$

then

$$z = z_0(s). \tag{0.2b}$$

The data is assumed to be noncharacteristic, i.e., by the chain rule we have

$$\dot{z}_0(s) = p_0(s)\dot{x}_0(s) + q_0(s)\dot{y}_0(s), \tag{0.3}$$

while

$$F(x_0(s), y_0(s), z_0(s), p_0(s), q_0(s)) = 0. \tag{0.4}$$

In order to solve (0.3) and (0.4) locally for smooth $p_0(s)$, $q_0(s)$, the implicit function theorem requires

$$\dot{y}_0 F_p(x_0, y_0, z_0, p_0, q_0) \neq \dot{x}_0 F_q(x_0, y_0, z_0, p_0, q_0). \tag{0.5}$$

This is the noncharacteristic criterion.

* Received by the editors October 1, 1992; accepted for publication (in revised form) December 16, 1992.

† Department of Mathematics, University of California, Los Angeles, California 90024. This research was supported by Army Research Office grant DAAL03-91-G0162 and National Science Foundation grant DMS 91-03104.

Given (0.5), one then generates characteristic curves via

$$\begin{aligned}\frac{dx}{dt} &= F_p,\\ \frac{dy}{dt} &= F_q,\\ \frac{dz}{dt} &= pF_p + qF_q,\\ \frac{dp}{dt} &= -F_x - pF_z,\\ \frac{dq}{dt} &= -F_y - qF_z.\end{aligned} \tag{0.6}$$

The initial data "propagates" along these curves and criterion (0.5) guarantees that we generate a smooth solution locally in time. However, in finite time characteristic curves generally intersect (caustics develop). Fourier Integral Operators (FIO) were developed in the sixties and seventies (see, e.g., [8]) to take care of the resulting multivaluedness (and to do a lot more, of course).

A classical example is the eiconal equation from geometrical optics:

$$p^2 + q^2 = 1. \tag{0.7}$$

If Γ is convex, the solution rapidly develops a caustic. Rather than continuing it as a multivalued solution à la FIO, we may use the recently developed notion of viscosity solutions [3], [4], [5] for Hamilton–Jacobi equations to continue the solution uniquely as a single-valued uniformly continuous function having "kinks"—i.e., jumps in the first derivative. For most real-world problems this is the appropriate class.

We shall propose an analytic and numerical method for solving (0.1), (0.2) when Γ is a compact set of closed curves dividing R^2 up into Ω and its complement Ω^c, neither of which needs to be compact. We call Ω the "interior" and Ω^c the "exterior."

This method generalizes easily to compact hypersurfaces dividing up R^n into an interior and exterior. In this paper we shall stick to R^2 for simplicity of exposition only.

The present work has three main antecedents. In [12] Osher and Sethian introduced the concept of a level set formulation to propagate curves and surfaces. The problem analyzed there was as follows. We wish to move a closed curve Γ normal to itself with normal velocity u_n. This velocity might be geometrically based; e.g., it might be a function of the curvature of Γ. The level set formulation easily treats self-intersections, topological changes, kinks, and higher space dimensions. Theoretical justification for this method (along with a great deal of other very important theory) came later in [2], [6], [7].

Briefly, one finds a function $\psi(x, y, t)$ so that at $t = 0$ we have

$$\psi(x,y,0) = 0 \Leftrightarrow (x,y) \in \Gamma, \tag{0.8a}$$

$$\psi(x,y,0) > 0 \quad \text{in } \Omega, \tag{0.8b}$$

$$\psi(x,y,0) < 0 \quad \text{in } \Omega^c, \tag{0.8c}$$

and $\psi(x, y, 0)$ is a uniformly continuous and monotonic strictly decreasing function of distance to Γ near Γ, which we call $\Gamma(0)$.

We require that $\Gamma(t)$ evolves so that

$$\psi(x,y,t) = 0 \Leftrightarrow (x,y) \in \Gamma(t). \tag{0.9a}$$

This means that for $(x(t), y(t)) \in \Gamma(t)$,

$$\frac{d}{dt}\psi(x(t), y(t), t) = 0, \tag{0.9b}$$

$$\psi_x x_t + \psi_y y_t + \psi_t = 0. \tag{0.9c}$$

Rearranging terms, we arrive at

$$\psi_t = -u_n\sqrt{\psi_x^2 + \psi_y^2}. \tag{0.9d}$$

At this point we imagine that u_n is defined throughout R^n, not just on $\Gamma(t)$, and that this is done in a natural way. Thus all level sets of ψ move according to this law.

If u_n is a given function of (x, y), then this is a Hamilton–Jacobi equation and we seek the viscosity solution [4], [5]. This has an interesting physical interpretation for flames. Sethian's entropy condition [15] follows for the viscosity solution (see [12] for the proof).

If u_n is the curvature of the level set, the equation becomes

$$\psi_t = \frac{\psi_{xx}\psi_y^2 - 2\psi_{xy}\psi_x\psi_y + \psi_{yy}\psi_x^2}{\psi_x^2 + \psi_y^2}. \tag{0.10}$$

Thus, we can define the motion of a square via its mean curvature using (0.10) and following the level set. Again, this was rigorously justified in [2], [7] for general curves modulo some unusual exceptions.

As a numerical device this approach has many advantages over tracking. We simply set up a fixed, Eulerian grid, solve (0.9) numerically, and let the plotter find the front. Self-intersections, kinks, topological changes, and multispace dimensions are treated routinely. Of course, we have to construct stable, accurate, and efficient methods for (0.9). See [12], [13] for a description of such methods.

The second antecedent is [10]. There the authors wished to solve a problem in computer vision. We are given $z(x, y)$, describing the surface of an object that is illuminated by an overhead light source at infinity. In the simplest model the intensity of light $I(x, y)$ is given by

$$I(x,y) - \frac{1}{\sqrt{1+p^2+q^2}} = 0 = F(x,y,z,p,q). \tag{0.11}$$

The shape-from-shading problem is: given $0 < I(x, y) \leqq 1$, find $u(x, y)$. This is a very well studied problem, but only recently in [11], [14] was the correct theory of viscosity solutions brought to consideration. In [10], the authors assumed that they were given a level surface of u, i.e., (0.2) for $z_0 \equiv 0$. What they proposed was to use the methods of [12] to propagate the level surface to generate the solution of (0.11). We now recognize this as a general method for solving (0.1) for Dirichlet data. We shall describe and justify it in the next section.

Finally, the third crucial antecedent came in Bruckstein [1]. There the author transformed the shape-from-shading problem into a level set propagation P.D.E. and realized the advantages of this formulation. The link with the propagation methods of [12] and the viscosity solution concept came later in [10] for this important problem.

1. Description and justification of the method. Let Γ be a compact set of disjoint closed curves in R^2, dividing R^2 up into an interior Ω and an exterior Ω^c. We wish to solve for $z = u(x,y)$,

$$F(x,y,z,p,q) = 0, \tag{1.1a}$$

with Dirichlet data on Γ written locally as

$$x = x_0(s), \tag{1.1b}$$

$$y = y_0(s), \tag{1.1c}$$

$$z = z_0(s), \tag{1.1d}$$

which is noncharacteristic for (1.1a).

We next assume that $z_0(s)$ can be continued into a set containing Γ as a function $w(x,y)$ so that the function $n(x,y)$ defined by

$$z(x,y) = w(x,y) + n(x,y) \tag{1.2}$$

is the unknown. This has the effect of changing F and setting $z_0 \equiv 0$ in (1.1d). Thus we have the zero level set of the solution to a simple related P.D.E. as boundary data. This new P.D.E. continues to be called F and the new unknown function is z.

The noncharacteristic criterion then becomes

$$p_0 F_p + q_0 F_q \neq 0 \tag{1.3}$$

on Γ.

Now we wish to construct a function of three variables $v(x,y,t)$, $t \geqq 0$ such that if

$$v(x,y,t) = 0, \quad \text{then} \tag{1.4a}$$

$$z = u(x,y) = t. \tag{1.4b}$$

Of course any such function will not be unique. However, all of them will satisfy on the level set (1.4):

$$\frac{\partial}{\partial x} v(x,y,u(x,y)) = 0 = v_x + v_t u_x, \tag{1.5a}$$

$$\frac{\partial}{\partial y} v(x,y,u(x,y)) = 0 = v_y + v_t u_y. \tag{1.5b}$$

Thus, at least formally on this level set,

$$F\left(x,y,t,\frac{-v_x}{v_t},\frac{-v_y}{v_t}\right) = 0 \quad \text{on } \{v = 0\}. \tag{1.6}$$

We shall choose $v(x,y,0)$ to be a uniformly continuous function vanishing only for (x,y) on Γ, $v > 0$ in Ω, $v < 0$ in Ω^c, and v is a strictly monotone function of distance to Γ near Γ.

The noncharacteristic criterion of (0.3) guarantees that we may invert (1.6) locally for v_t near Γ. To devise a numerical algorithm based on time evolution we need the following assumption.

Assumption 1. An explicit inversion formula exists for (1.6) near Γ so that the formula

$$v_t + H(x, y, t, v_x, v_y) = 0 \tag{1.7}$$

with $H > 0$ near Γ implies (1.6) near $(x, y) \in \Gamma$, $t = 0$.

We note that H must be homogeneous of degree one in v_x, v_y.

Some geometric analysis of Assumption 1 is in order. We wish to solve (1.6) for v_t. By the implicit function theorem this is valid if (1.3) is valid near Γ, i.e.,

$$pF_p + qF_q \neq 0.$$

As pointed out by Evans, this says that the zero level set (in (p, q) space) of F is star shaped.

Another interesting observation of Evans concerns the link with steady multidimensional conservation laws. Equation (1.7) is an ordinary Hamilton–Jacobi equation, while (1.1) might be a conservation law admitting shock solutions. This may provide a connection between conservation laws and Hamilton–Jacobi equations in multidimensions.

We now have our analytical method for solving (1.1), kinks and all. (Numerical methods may be easily constructed using the results of [12], [13].)

We solve (1.7) on all of R^n (we really only need to do this near $\Gamma(t)$) with uniformly continuous initial data.

$$v(x, y, 0) = v_0(x, y), \tag{1.8a}$$

with

$$v_0(x, y) = 0 \quad \text{if and only if } (x, y) \in \Gamma, \tag{1.8b}$$

$$v_0(x, y) > 0 \quad \text{if and only if } (x, y) \in \Omega, \tag{1.8c}$$

$$v_0(x, y) < 0 \quad \text{if and only if } (x, y) \in \Omega^c. \tag{1.8d}$$

Then, to compute $u(x, y)$ for $(x, y) \in \Omega$ we calculate the level sets via the relation

$$v(x, y, t) = 0 \Leftrightarrow t = u(x, y). \tag{1.9}$$

This allows us to generate $u(x, y)$ by building it up through this level set formulation.

It is clear from the classical method of characteristics (see, e.g., [9]) that if Γ is a smooth curve and F and H are smooth functions near Γ, then the solution to (1.1) is locally (near Γ) the same as (1.9) for $t > 0$ and small. We now claim that the level set generated function (1.9) is a viscosity solution to

$$-1 + H(x, y, u, u_x, u_y) = 0 \tag{1.10}$$

if v is the viscosity solution to (1.7), (1.8).

We now recall the definition of viscosity solution; see, e.g., [3].

DEFINITION 1.1. Let $\psi \in C^2$ near $(\bar{x}, \bar{y}, \bar{t})$. Suppose $v - \psi$ has a local minimum (maximum) at $(\bar{x}, \bar{y}, \bar{t})$. Then v is a viscosity supersolution (subsolution) of (1.7) at this point if

$$\psi_t + H(x, y, v, \psi_x, \psi_y) \geqq 0 (\leqq 0) \text{at } (\bar{x}, \bar{y}, \bar{t}) \tag{1.11}$$

for all such ψ.

DEFINITION 1.2. v is a viscosity solution at this point if it is both a viscosity sub and supersolution.

The fact that $H > 0$ indicates that v is strictly decreasing in t near this point. In fact, if $v - \psi$ has a local maximum there, then

$$\psi_t < -H,$$

which means that ψ is strictly decreasing there.

We take ψ so that

$$v(x, y, t) \leqq \psi(x, y, t) \quad \text{near } (\bar{x}, \bar{y}, \bar{t}) \tag{1.12a}$$

and

$$v(\bar{x}, \bar{y}, \bar{t}) = \psi(\bar{x}, \bar{y}, \bar{t}). \tag{1.12b}$$

Then, for t satisfying $\bar{t} < t < \bar{t} + \varepsilon$ for $\varepsilon > 0$ small,

$$\begin{aligned} v(\bar{x}, \bar{y}, t) \; - \; & v(\bar{x}, \bar{y}, \bar{t}) \leqq \psi(\bar{x}, \bar{y}, t) - \psi(\bar{x}, \bar{y}, \bar{t}) \\ & = \psi_t(\bar{x}, \bar{y}, \bar{t})(t - \bar{t}) \quad (\text{where } \bar{t} < \bar{t} < t) \\ & < -\bar{H}(t - \bar{t}). \end{aligned} \tag{1.13}$$

Thus v is uniformly strictly decreasing for $t > \bar{t}$. This is true for all such $\bar{t}$ and t in any neighborhood in which $v(x, y, t)$ is a viscosity solution. Thus there exists an increasing uniformly continuous inverse function h such that

$$h(v(x, y, t)) = u(x, y) - t. \tag{1.14}$$

What remains to be shown is that u is a viscosity solution to

$$-1 + H(x, y, u, u_x, u_y) = 0. \tag{1.15}$$

This follows directly from [2, Thm. 5.2] under the hypothesis that H is independent of u. Thus we make that assumption for theoretical purposes only and conclude. (This key theorem of [2] was motivated by problems involving motion of level sets such as those described in [12].)

2. Examples. Given the shape-from-shading problem described above, with a remote generally nonoverhead light source whose direction cosines are $(\alpha, \beta, -\gamma)$ for $\gamma > 0$, with respect to the normal to the surface $z = u(x, y)$, we wish to solve

$$I(x, y)\sqrt{1 + u_x^2 + u_y^2} - \alpha u_x - \beta u_y - \gamma = 0 \tag{2.1a}$$

with

$$u = 0 \quad \text{on } \Gamma = \partial\Omega. \tag{2.1b}$$

The noncharacteristic criterion is satisfied if $\gamma\sqrt{1+p^2+q^2} \neq I$. Equation (1.6) becomes in this case

$$I(x,y)\sqrt{1+\frac{v_x^2}{v_t^2}+\frac{v_y^2}{v_t^2}}+\frac{\alpha v_x}{v_t}+\frac{\beta v_y}{v_t}-\gamma=0. \tag{2.2}$$

This can be inverted to obtain our version of (1.7):

$$v_t+\frac{(\mathrm{sign}(I^2-\gamma^2))\gamma(\alpha v_x+\beta v_y)+I\sqrt{v_x^2(-\beta^2+1-I^2)+v_y^2(-\alpha^2+1-I^2)+2\alpha\beta v_x v_y}}{|\gamma^2-I^2|}=0. \tag{2.3}$$

Note that if, for example, $\gamma = 1 \Leftrightarrow \alpha = \beta = 0$, the resulting overhead formula

$$v_t+\frac{I}{\sqrt{1-I^2}}\sqrt{v_x^2+v_y^2}=0 \tag{2.4}$$

gives difficulties near $I = 1$. This is inherent in the problem [11], [14]. In the general case the method has problems because of a possibly negative quantity under the square root sign, unless $I \leqq \gamma$. If this inequality fails we must require that the gradients satisfy

$$v_x^2(1-I^2-\beta^2)+v_y^2(1-I^2-\alpha^2)+2\alpha\beta v_x v_y \geqq 0 \tag{2.5a}$$

if

$$I>\gamma. \tag{2.5b}$$

This is, of course, required at the zero level set of $v(x,y,t)$ from (2.1a) using (2.3) and the related (1.5a, b), but it does present some numerical difficulties.

Example 2. Control-optimal cost determination:

$$-(\sin y)u_x+(\sin x)u_y+|u_y|-\tfrac{1}{2}\sin^2 y-(1-\cos x)=0, \tag{2.6a}$$

$$u=0 \quad \text{on } \Gamma, \text{ which is noncharacteristic, which means} \tag{2.6b}$$

$$\tfrac{1}{2}\sin^2 y+(1-\cos x)\neq 0 \quad \text{on } \Gamma. \tag{2.6c}$$

We are led to

$$v_t+\frac{|v_y|+(\sin x)v_y-(\sin y)v_x}{1/2\sin^2 y+(1-\cos x)}=0, \tag{2.7}$$

and the quantity $H(x,y,u_x,u_y)$ defined above is assumed to be strictly positive near Γ. (Of course, if it is strictly negative, everything works with a different initialization, merely reversing the inequalities in (1.8c, d).)

Acknowledgments. We would like to thank Professor A. M. Bruckstein for making us aware of his work in this area and Professor L. C. Evans for informing us of the existence of Theorem 5.2 in [2] and for some very constructive comments on a first draft of the paper.

REFERENCES

[1] A. M. Bruckstein, *On shape from shading*, Comput. Vision Graphics Image Process, 44 (1988), pp. 139–154.

[2] Y. G. Chen, Y. Giga, and S. Goto, *Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations*, J. Differential Geom., 33 (1991), pp. 749–786.

[3] M. G. Crandall, H. Ishii, and P. L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Amer. Math. Soc. Bull., 27 (1992), pp. 1–67.

[4] M. G. Crandall and P. L. Lions, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 227 (1983), pp. 1–42.

[5] ———, *On existence and uniqueness of solutions of Hamilton–Jacobi equations*, Nonlinear Anal. Theory Meth. Appl., 10 (1986), pp. 353–370.

[6] L. C. Evans, M. Soner, and P. E. Souganidis, *The Allen–Cahn equation and generalized motion by mean curvature*, preprint.

[7] L. C. Evans and J. Spruck, *Motion of level sets by mean curvature* I, J. Differential Geom., 33 (1991), pp. 635–681.

[8] L. Hormander, *The calculus of Fourier Integral Operations*, in Ann. Math. Stud. 70, Prospects in Mathematics, Princeton University Press, Princeton, NJ, 1971, pp. 35–57.

[9] F. John, *Partial Differential Equations*, fourth ed., Springer-Verlag, New York, 1982.

[10] R. Kimmel and A. M. Bruckstein, *Shape from Shading via Level Sets*, Technion (Israel) Computer Science Dept. Report, CIS #9209, 1992.

[11] P. L. Lions, E. Rouy, and A. Tourin, *Shape-from-shading, viscosity solutions, and edges*, Numer. Math., to appear.

[12] S. J. Osher and J. A. Sethian, *Fronts propagating with curvature dependent speed: algorithms based on Hamilton–Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.

[13] S. Osher and C.-W. Shu, *High-order essentially nonoscillatory schemes for Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 28 (1991), pp. 907–922.

[14] E. Rouy and A. Tourin, *A viscosity solution approach to shape from shading*, SIAM J. Numer. Anal., 29 (1992), pp. 867–884.

[15] J. A. Sethian, *Curvature and the Evolution of Fronts*, Comm. Math. Phys., 10 (1985), pp. 487–499.

SIAM J. MATH. ANAL.
Vol. 24, No. 5, pp. 1153–1178, September 1993

003

THREE-DIMENSIONAL STEADY WATER WAVES GENERATED BY PARTIALLY LOCALIZED PRESSURE DISTURBANCES*

TIEN-YU SUN†

Abstract. The author constructs a class of three-dimensional exact steady water waves resulting from a partially localized pressure disturbance on the free surface. The term *partially localized* means that the pressure disturbances are periodic in the direction of the flow and are decaying rapidly in the transverse direction. The resulting exact steady flows exhibit symmetric doubly periodic wave patterns at infinity on either side of the pressure disturbance. The surface tension effect is taken into account, and this enables the author to use the Implicit Function Theorem in his construction.

Key words. water waves, free boundary problem, inviscid incompressible flows

AMS subject classifications. 35R35, 76B15

1. Introduction. We are concerned with steady wave motions of an infinite ocean of water, bounded above by a free surface and below by a flat bottom. The problem is to determine the velocity field and the free surface as an exact steady solution of the equations of water waves; see (2.1)–(2.5). Earlier work on exact steady water waves used complex formulation of the problem, and hence the results were limited to two-dimensional flows. See Wehausen and Laitone [7] for a detailed reference. Methods applicable to three-dimensional flows were considered by Beale [1] and Hewgill, Reeder, and Shinbrot [2]. In [5], with no pressure disturbance on the free surface, Reeder and Shinbrot proved the existence of symmetric doubly periodic solutions of (2.1)–(2.5), treating them as nonlinear interactions of two periodic plane waves. The surface tension effect is crucial in their construction. In contrast to the two-dimensional periodic waves, the existence of three-dimensional symmetric doubly periodic solutions becomes a small divisor problem when the surface tension effect is omitted. See Plotnikov [4].

In this paper, the method developed by Beale in the two-dimensional setting is applied to the three-dimensional problem. By applying a specific class of pressure disturbances on the free surface, we show that a class of three-dimensional exact steady solutions can be constructed as in [1]. It is well known that steady flows of water resulting from pressure disturbances on the free surface may have various wave patterns at infinity. For two-dimensional flows, it has been shown that the steady wave generated by a localized pressure disturbance has periodic wave patterns at infinity. When the surface tension effect is included, capillary waves appear upstream and gravity waves appear downstream. See Lighthill [3] and Whitham [8] for the linear theory and Beale [1] for the nonlinear result. The wave patterns at infinity remain a major concern in constructing three-dimensional steady water waves. Further complication arises from the fact that a three-dimensional linear steady wave may exhibit a complicated wave pattern like the wake after a ship. Again see Lighthill [3] and Whitham [8] for the linear theory of ship waves.

The main result of this paper can be stated as follows. We assume that the pressure disturbances applied on the free surface are partially localized and are small in amplitude. By partially localized, we mean the pressure disturbances are periodic

* Received by the editors July 15, 1991; accepted for publication (in revised form) November 30, 1992.

† Department of Mathematics, Chung-Yuan Christian University, Chung-Li, Taiwan 32023 Republic of China.

in the direction of the flow and are decaying rapidly in the transverse direction. We seek exact solutions which are small perturbations of the uniform horizontal flow with constant flow speed U_0 and constant depth h. When the flow speed U_0 and the period of the pressure disturbance are within a certain range, we show that there exists a two-parameter family of three-dimensional exact steady waves which exhibit symmetric doubly periodic wave patterns at infinity on either side of the pressure disturbance. The symmetric doubly periodic waves appearing are those constructed by Reeder and Shinbrot in [5]. The two parameters involved are a phase shift at infinity on each side of the partially localized pressure disturbance. If the pressure disturbance is even in the transverse direction, the resulting solutions can be regarded as steady waves traveling along a vertical wall. As in [5], the surface tension effect is important in our construction.

In §2, the problem is formulated in detail. In the problem, we assume that a partially localized pressure disturbance with wave number k_1 in the direction of flow is applied on the free surface. In §3, symmetric doubly periodic waves with flow speed U_0 and wave number k_1 in the direction of the flow are constructed. In comparison to [5], §3 provides an alternative construction of the symmetric doubly periodic waves which uses a different set of relevant parameters. In §4, exact steady flows resulting from a partially localized pressure disturbances on the free surface are constructed.

2. Preliminaries. We begin by describing the three-dimensional equations of water waves, with the surface tension effect taken into account on the free surface. Suppose a coordinate system (X, Y, Z) is chosen so that the X, Y directions are horizontal and the Z direction points upward. Let (U, V, W) and $Z = S(X, Y)$ be the unknown velocity field and free surface, respectively. Assuming the ocean has a flat bottom, the fluid occupies the domain $\{(X, Y, Z) : 0 < Z < S(X, Y)\}$. Now the equations of water waves can be stated as follows:

$$U_X + V_Y + W_Z = 0, \tag{2.1}$$

$$\begin{aligned} W_Y - V_Z &= 0, \\ U_Z - W_X &= 0, \\ V_X - U_Y &= 0, \end{aligned} \tag{2.2}$$

$$W = 0 \text{ on } Z = 0, \tag{2.3}$$

$$W - US_X - VS_Y = 0 \text{ on } Z = S(X, Y), \tag{2.4}$$

$$gS - \beta_0 D \cdot \frac{DS}{\sqrt{1 + |DS|^2}} + \frac{1}{2}(U^2 + V^2 + W^2) + \rho_0{}^{-1}(P - P_0) = C \quad \text{on } Z = S(X, Y), \tag{2.5}$$

where $D = (\partial_X, \partial_Y)$. Here P is the pressure on the free surface, P_0 is the atmospheric pressure, and ρ_0 is the constant density of the fluid. The parameters g and β_0 are the gravitational acceleration and the surface tension coefficient, respectively. Equations (2.1) and (2.2) are due to the incompressibility and irrotationality of the fluid; (2.3) and (2.4) are the streamline conditions on the bottom and top surfaces; (2.5) is Bernoulli's equation.

When $P = P_0$, the uniform horizontal flow with $S(X,Y) = h > 0$, $U = U_0$ and $V = W = 0$ is a trivial solution of the above boundary value problem. Hence, when $P - P_0$ is small, we seek exact solutions of (2.1) – (2.5) which are small perturbations of the uniform horizontal flow. On the free surface, we suppose $P - P_0 = \epsilon\, \rho_0\, ghp(X,Y)$, where

(2.6) p is periodic in X and is decaying rapidly in the Y direction.

The factor $\rho_0 gh$ is inserted for convenience. We can assume that the velocity (U, V, W) and the free surface S have the form

$$\begin{aligned} U &= U_0\,(1 + \epsilon\, u), \\ V &= \epsilon\, U_0\, v, \quad W = \epsilon\, U_0\, w, \\ S &= h\,(1 + \epsilon\, \eta). \end{aligned} \tag{2.7}$$

By stretching the vertical coordinate, we can transform the fluid domain to a fixed horizontal slab. We set

$$\begin{aligned} x &= X/h, \qquad y = Y/h, \\ z &= Z/h(1 + \epsilon\, \eta). \end{aligned}$$

Now the new fluid domain is $\{(x, y, z) : 0 < z < 1\}$; and (2.1)–(2.5) are transformed to

$$\{\,(1 + \epsilon\,\eta)\,u_x - \epsilon\, z\, \eta_x\, u_z\,\} + \{\,(1 + \epsilon\,\eta)\,v_y - \epsilon\, z\, \eta_y\, v_z\,\} + w_z = 0, \tag{2.8}$$

$$\begin{aligned} w_y - \sigma\, v_z - \epsilon\, z\, \sigma\, \eta_y\, w_z &= 0, \\ \sigma\, u_z - w_x + \epsilon\, z\, \sigma\, \eta_x\, w_z &= 0, \\ v_x - u_y - \epsilon\, z\, \sigma\, \eta_x\, v_z + \epsilon\, z\, \sigma\, \eta_y\, u_z &= 0, \end{aligned} \tag{2.9}$$

$$w = 0 \quad \text{on } z = 0, \tag{2.10}$$

$$w - \epsilon\, u\, \eta_x - \epsilon\, v\, \eta_y = \eta_x \quad \text{on } z = 1, \tag{2.11}$$

$$\eta - \beta \nabla \cdot \frac{\nabla \eta}{\sqrt{1 + \epsilon^2\, |\, \nabla \eta\, |^2}} + \gamma\, u + \frac{\epsilon}{2}\, \gamma\, (\, u^2 + v^2 + w^2\,) + p = 0 \quad \text{on } z = 1, \tag{2.12}$$

where $\nabla = (\,\partial_x,\, \partial_y\,)$ and $\sigma = (1 + \epsilon\, \eta)^{-1}$. The parameters

$$\beta = \beta_0\, /\, gh^2, \qquad \gamma = {U_0}^2\, /\, gh$$

are the dimensionless coefficient of surface tension and the Froude number, respectively.

The free surface is now removed at the price of introducing nonlinear terms into the interior equations. For ϵ sufficiently small, we can treat η as already known and solve (2.8)–(2.11) for (u, v, w). This enables us to regard (u, v, w) as a function of η, and we are led to solve (2.12) as a functional equation of η. Linear approximations can be found by solving (2.8)–(2.12) with $\epsilon = 0$. Based on the solvability of the linear problem, exact solutions are constructed using the implicit function theorem.

Now we introduce the function spaces that are used in the subsequent sections. Let $\ell_1 = \pi/k_1$ and

$$\begin{aligned}\Omega &= \{(x,y,z) : -\ell_1 < x < \ell_1,\, 0 < z < 1\},\\ \Gamma &= \{(x,y) : -\ell_1 < x < \ell_1\}.\end{aligned}$$

For a nonnegative integer s, let $H^s(\Omega)$ and $H^s(\Gamma)$ be the Sobolev spaces of functions that are periodic in x with period $2\ell_1$, with derivatives up to order s in L^2. A subscript e or o is used to indicate the subspaces of functions that are even or odd in x. In particular, for $v \in H^s_e(\Gamma)$, v can be written as

$$v = \sum_{m=0}^{+\infty} v_m(y) \cos mk_1x,$$

and its H^s norm is given by

$$|\, v\,|_s = \left\{ \sum_{m=0}^{+\infty} \int_{-\infty}^{+\infty} (1 + m^2{k_1}^2 + \xi^2)^s \,|\, \widehat{v_m}(\xi)\,|^2 \, d\xi \right\}^{1/2}.$$

Here $\widehat{\ }$ is the Fourier transform in y with ξ as the dual variable. When $s > 0$ is nonintegral, these spaces are defined according to the usual generalization. Given $\rho > 0$, ${}_\rho H^s(\Omega)$ and ${}_\rho H^s(\Gamma)$ represent subspaces of $H^s(\Omega)$ and $H^s(\Gamma)$ consisting of functions v which decay rapidly in the y direction such that $(\cosh \rho y) v \in H^s$, with norms $|\, v\,|_{s,\rho} = |\, (\cosh \rho y)\, v\,|_s$. Again, a subscript e or o is used to indicate subspaces of ${}_\rho H^s$ consisting of functions which are even or odd in x.

As for functions which are doubly periodic in x and y with periods $2\ell_1 = 2\pi/k_1$ and $2\ell_2 = 2\pi/k_2$, let

$$\begin{aligned}\widetilde{\Omega} &= \{(x,y,z) : -\ell_1 < x < \ell_1,\, -\ell_2 < y < \ell_2,\, 0 < z < 1\},\\ \widetilde{\Gamma} &= \{(x,y) : -\ell_1 < x < \ell_1,\, -\ell_2 < y < \ell_2\}.\end{aligned}$$

$H^s(\widetilde{\Omega})$ and $H^s(\widetilde{\Gamma})$ represent Sobolev spaces of functions doubly periodic in x and y, whose derivatives up to order s are in L^2, provided $s \geq 0$ is an integer, or the usual generalization when s is nonintegral but > 0. A subscript (e,e), (e,o), (o,e), or (o,o) is used to indicate subspaces with specific symmetries in x and y. For example, $H^s_{e,o}(\widetilde{\Omega})$ are the subspaces of $H^s(\widetilde{\Omega})$ consisting of functions that are even in x and odd in y.

3. Symmetric doubly periodic waves. The problem in the next section is to construct exact steady waves traveling with flow speed U_0 in the x direction, generated by a partially localized pressure disturbance with wave number k_1 in the same direction. In this section, we construct a family of exact doubly periodic solutions of (2.8)–(2.12) with $p = 0$. As we will see, these exact solutions are the symmetric doubly periodic waves that the solutions of our problem tend to as $y \to \pm\infty$. In [5], Reeder and Shinbrot proved the existence of symmetric doubly periodic water waves by regarding them as nonlinear interactions of incoming periodic plane waves with their reflections off a vertical wall. Incident wavelengths and incident angles within a specific set were used as the parameters of the problem. In this regard, the present section provides an alternative construction which uses a different set of relevant parameters.

In what follows, we will assume that the symmetric doubly periodic waves considered have flow speed U_0 and wave number k_1 in the x direction. We first consider (2.8)–(2.12) with $\epsilon = 0$ and assume the linear approximation considered has η given by

$$\eta(x,y) = \cos k_1 x \, \cos k_2 y. \tag{3.1}$$

Here the wave number k_2 is yet to be determined. With η given by (3.1), it is easy to solve (2.8)–(2.11) and obtain

$$\begin{aligned} u &= \frac{-k_1{}^2}{\mid k \mid} \frac{\cosh \mid k \mid z}{\sinh \mid k \mid} \cos k_1 x \, \cos k_2 y, \\ v &= \frac{k_1 k_2}{\mid k \mid} \frac{\cosh \mid k \mid z}{\sinh \mid k \mid} \sin k_1 x \, \sin k_2 y, \\ w &= -k_1 \frac{\sinh \mid k \mid z}{\sinh \mid k \mid} \sin k_1 x \, \cos k_2 y, \end{aligned} \tag{3.2}$$

where $k = (k_1, k_2)$ and $\mid k \mid = (k_1{}^2 + k_2{}^2)^{1/2}$. Consequently,

$$\eta - \beta \nabla^2 \eta + \gamma u = b(k_1, k_2) \cos k_1 x \, \cos k_2 y.$$

Here the function $b(\cdot,\cdot)$ is given by

$$b(\tau_1, \tau_2) = 1 + \beta \mid \tau \mid^2 - \gamma \frac{\tau_1{}^2}{\mid \tau \mid} \coth \mid \tau \mid, \tag{3.3}$$

where $\tau = (\tau_1, \tau_2)$ and $\mid \tau \mid = (\tau_1{}^2 + \tau_2{}^2)^{1/2}$. Hence, in order for η and (u, v, w) given by (3.1) and (3.2) to satisfy (2.12), we need wave number k_2 to satisfy $b(k_1, k_2) = 0$.

Now, for each fixed parameter $\beta > 0$, define

$$\gamma_0(\beta) = \min_{\xi \in \mathbf{R}} \frac{1 + \beta \xi^2}{\xi \coth \xi}.$$

It was pointed out in Lemma 2 in [1] that $\gamma_0(\beta) < 1$, provided $\beta < \frac{1}{3}$. Moreover, when $\beta < 0.02$ and the flow speed U_0 is within a range so that the Froude number γ satisfies $\gamma_0(\beta) < \gamma < 1$, the equation

$$1 + \beta \xi^2 - \gamma \xi \coth \xi = 0$$

has two simple positive roots, k_G, k_T, where $0 < k_G < k_T$. Note that, from (3.3), we have $b(k_G, 0) = b(k_T, 0) = 0$. The following lemma is used to determine the wave number k_2 in the transverse direction.

LEMMA 3.1. *Suppose β and γ are fixed parameters with $0 < \beta < 0.02$ and $\gamma_0(\beta) < \gamma < 1$. Let $b(\cdot,\cdot)$ be the function defined in (3.3). For $\tau_1 > 0$ so that $k_G < \tau_1 < k_T$, the equation $b(\tau_1, \xi) = 0$ has a unique positive root $\xi = k_2$ which is simple. When $0 < \tau_1 < k_G$ or $k_T < \tau_1$, the equation $b(\tau_1, \xi) = 0$ has no real roots.*

Proof. See Appendix. □

In this paper, we will always assume that the water is deep enough such that $\beta < 0.02$. When $\gamma_0(\beta) < \gamma < 1$ and $k_G < k_1 < k_T$, with wave number k_2 determined by solving $b(k_1,\xi) = 0$, η and (u,v,w) given by (3.1) and (3.2) provide us a linear approximation of symmetric doubly periodic waves.

The wave number k_2 determined by the flow speed U_0 and wave number k_1 provide us a first approximation of the wave number in the transverse direction. For fixed U_0 and k_1, an exact solution of (2.8)–(2.12) based on linear approximation (3.1), (3.2) should have wave number in the y direction fairly close to k_2 when ϵ is sufficiently small. We can assume that the exact wave number in the y direction is $\theta(\epsilon)k_2$. Here we suppose $\theta = \theta(\epsilon)$ is close to 1 for ϵ sufficiently small. Now we stretch the y coordinate axis by setting $\tilde{y} = \theta y$ and replace the (x, y, z) coordinate system by $(x, \tilde{y}, z)$ instead. As a result, (2.8)–(2.12) are transformed to

$$\{(1+\epsilon\eta)\,u_x - \epsilon\, z\,\eta_x\, u_z\} + \theta\,\{(1+\epsilon\eta)\,v_{\tilde{y}} - \epsilon\, z\,\eta_{\tilde{y}}\, v_z\} + w_z = 0, \tag{3.4}$$

$$\begin{aligned} \theta\, w_{\tilde{y}} - \sigma\, v_z - \epsilon\,\theta\, z\,\sigma\,\eta_{\tilde{y}}\, w_z &= 0,\\ \sigma\, u_z - w_z + \epsilon\, z\,\sigma\,\eta_x\, w_z &= 0,\\ v_x - \theta\, u_{\tilde{y}} - \epsilon\, z\,\sigma\,\eta_x\, v_z + \epsilon\,\theta\, z\,\sigma\eta_{\tilde{y}}\, u_z &= 0, \end{aligned} \tag{3.5}$$

$$w = 0 \quad \text{on } z = 0, \tag{3.6}$$

$$w - \epsilon\, u\,\eta_x - \epsilon\,\theta\, v\,\eta_{\tilde{y}} = \eta_x \quad \text{on } z = 1, \tag{3.7}$$

$$\eta - \beta\tilde{\nabla}\cdot\frac{\tilde{\nabla}\eta}{[1+\epsilon^2\,|\,\tilde{\nabla}\eta\,|^2]^{1/2}} + \gamma\, u + \frac{\epsilon}{2}\,(u^2+v^2+w^2) = 0 \quad \text{on } z = 1. \tag{3.8}$$

Here $\tilde{\nabla} = (\,\partial_x,\ \theta\,\partial_{\tilde{y}}\,)$ and $\sigma = (1+\epsilon\eta)^{-1}$. Thus, by introducing the stretching factor θ as an unknown, the doubly periodic waves sought now have a fixed period $2\ell_2 = 2\pi/k_2$ in the $\tilde{y}$ direction. In comparison to the construction in [5], the symmetric doubly periodic waves considered there have both of the periods determined by the incident wavelength and the incident angle. Hence no stretching of the coordinate axis was needed to determine the domain occupied by the symmetric doubly periodic waves. However, the Froude number had to be determined as part of the solution.

A careful examination of (3.4)–(3.8) suggests that we can assume the exact solution has the same symmetries in x and $\tilde{y}$ as in (3.1) and (3.2). In particular, we seek η and (u,v,w) so that η, u are even in x and v, w are odd in x. An immediate consequence is that $V = \epsilon\, U_0\, v$ is identically zero along $x = 0$. Generally speaking, we cannot expect (2.1)–(2.5) to have a unique exact doubly periodic solution unless some kind of normalization is imposed. For this reason, we suppose

$$\int_0^{S(0,Y)} U(0,Y,Z)\,dZ = U_0\, h. \tag{3.9}$$

In terms of the new coordinates, (3.9) can be written as

$$(1+\epsilon\eta)\int_0^1 u\,dz + \eta = 0 \quad \text{on } x = 0. \tag{3.10}$$

Integrating (3.10) with respect to $\tilde{y}$, we obtain

$$\int_{-\ell_2}^{\ell_2}\int_0^1 (1+\epsilon\eta)\,u\,dz d\tilde{y} = -\int_{-\ell_2}^{\ell_2} \eta\, d\tilde{y} \quad \text{on } x = 0. \tag{3.11}$$

In what follows, we will solve (3.4)–(3.8) with (3.11) as the normalization. When $\epsilon = 0$, this modified system has $\theta = 1$, and η, (u, v, w) given by (3.1) and (3.2) as solution.

We start the construction by finding (u, v, w) that satisfies (3.4)–(3.7) and (3.11) for prescribed η, θ, and ϵ. We suppose that $\eta \in H^{s+3/2}_{e,e}(\tilde{\Gamma})$, where s is an integer $> \frac{3}{2}$. Let

$$Y = \{\, (u, v, w) \in H^{s+1}_{e,e}(\tilde{\Omega}) \times H^{s+1}_{o,o}(\tilde{\Omega}) \times H^{s+1}_{o,e}(\tilde{\Omega}) : w(x, \tilde{y}, 0) = 0 \,\}$$

and

$$Z = H^{s}_{o,e}(\tilde{\Omega}) \times H^{s}_{o,o}(\tilde{\Omega}) \times H^{s}_{e,e}(\tilde{\Omega}) \times H^{s}_{e,o}(\tilde{\Omega}) \times H^{s+1/2}_{o,e}(\tilde{\Gamma}) \times \mathbf{R}.$$

We can rewrite (3.4), (3.5), (3.7), and (3.11) as

$$L(\epsilon\,\eta, \theta)(u, v, w) = \left(0, \vec{0}, \eta_x, -\int_{-\ell_2}^{\ell_2} \eta(0, \tilde{y})\, d\tilde{y} \right), \tag{3.12}$$

where $L(\epsilon\,\eta, \theta)$ is a bounded linear operator from Y to Z, depending smoothly on η, θ, and ϵ. To examine the range of $L(\epsilon\,\eta, \theta)$, suppose $\vec{g} = (g_1, g_2, g_3)$ and $(f, \vec{g}, q, c) \in Z$ such that

$$L(\epsilon\,\eta, \theta)(u, v, w) = (f, \vec{g}, q, c) \tag{3.13}$$

for some (u, v, w) in Y. Note that, from (3.13), we see that $\vec{g}$ is a curl in the (X, Y, Z) coordinates and thus its divergence with respect to (X, Y, Z) has to be zero. In terms of the new coordinates, $\vec{g}$ satisfies

$$\{\, (1 + \epsilon\,\eta)\, g_{1,x} - \epsilon\, z\, \eta_x\, g_{1,z} \,\} + \theta \{\, (1 + \epsilon\,\eta)\, g_{2,\tilde{y}} - \epsilon\, z\, \eta_{\tilde{y}}\, g_{2,z} \,\} + g_{3,z} = 0. \tag{3.14}$$

This indicates that the range of $L(\epsilon\,\eta, \theta)$ is a subspace of Z which varies with η, θ, and ϵ. For each $\vec{g}$, define $\vec{\tilde{g}} = (\tilde{g}_1, \tilde{g}_2, \tilde{g}_3)$ by

$$\begin{gathered} \tilde{g}_1 = (1 + \epsilon\,\eta)\, g_1, \qquad \tilde{g}_2 = \theta\, (1 + \epsilon\,\eta)\, g_2, \\ \tilde{g}_3 = g_3 - \epsilon\, z\, \eta_x\, g_1 - \epsilon\, \theta\, z\, \eta_{\tilde{y}}\, g_2, \end{gathered} \tag{3.15}$$

and let $J(\epsilon\,\eta, \theta) : Z \to Z$ be the linear operator which maps $(f, \vec{g}, q, c)$ to $(f, \vec{\tilde{g}}, q, c)$. Let Z_0 be the subspace of Z consisting of $(f, \vec{g}, q, c)$ such that

$$g_{1,x} + g_{2,\tilde{y}} + g_{3,z} = 0. \tag{3.16}$$

Now if $(f, \vec{g}, q, c) \in Z$ is in the range of $L(\epsilon\,\eta, \theta)$, then (3.14) implies that $\vec{\tilde{g}}$ defined by (3.15) satisfies

$$\tilde{g}_{1,x} + \tilde{g}_{2,\tilde{y}} + \tilde{g}_{3,z} = 0.$$

That is to say, $J(\epsilon\,\eta, \theta)$ maps the range of $L(\epsilon\,\eta, \theta)$ into Z_0, which is independent of η, θ, and ϵ. It is clear that operator J is bounded and is depending smoothly on η, θ, and ϵ. Moreover, $J(0, 1)$ is the identity operator. Hence, for η in a bounded set in $H^{s+3/2}_{e,e}(\tilde{\Gamma})$ and θ close to 1, $J(\epsilon\,\eta, \theta)$ is invertible when ϵ is sufficiently small. For this reason, since

$$J(\epsilon\,\eta, \theta) \left(0,\, \vec{0},\, \eta_x,\, -\int_{-\ell_2}^{\ell_2} \eta(0, \tilde{y})\, d\tilde{y} \right) = \left(0,\, \vec{0},\, \eta_x,\, -\int_{-\ell_2}^{\ell_2} \eta(0, \tilde{y})\, d\tilde{y} \right),$$

(3.12) and

$$(JL)(\epsilon\eta,\theta)(u,v,w) = \left(0,\ \vec{0},\ \eta_x,\ -\int_{-\ell_2}^{\ell_2} \eta(0,\tilde{y})\,d\tilde{y}\right)$$

share the same solution. In the following lemma, we show that $JL(0,1) = L(0,1)$, as an operator from Y to Z_0, is invertible.

LEMMA 3.2. *Given $(f,\vec{g},q,c) \in Z_0$, there is a unique solution in Y of the problem*

$$u_x + v_{\tilde{y}} + w_z = f, \tag{3.17}$$

$$w_{\tilde{y}} - v_z = g_1, \tag{3.18}$$

$$u_z - w_x = g_2, \tag{3.19}$$

$$v_x - u_{\tilde{y}} = g_3, \tag{3.20}$$

$$w = q \quad on\ z = 1, \tag{3.21}$$

$$w = 0 \quad on\ z = 0, \tag{3.22}$$

$$\int_{-\ell_2}^{\ell_2}\int_0^1 u(0,\tilde{y},z)\,dz d\tilde{y} = c. \tag{3.23}$$

The solution (u,v,w) satisfies

$$|\,u\,|_{s+1} + |\,v\,|_{s+1} + |\,w\,|_{s+1} \le C\left(|\,f\,|_s + \sum_{i=1}^{3} |\,g_i\,|_s + |\,q\,|_{s+1/2} + |\,c\,|\right). \tag{3.24}$$

Since the proof is similar to that of Lemma 5 in [1], we will summarize it as follows. First, we consider the case with $q = 0$. Assume that the solution (u,v,w) is known for the time being. Let

$$l_n(z) = \sqrt{2}\cos n\pi z, \qquad m_n(z) = \sqrt{2}\sin n\pi z$$

for $n \ge 1$, and let $l_0(z) = 1$. To expand u, v, and w in eigenfunctions in z, we define

$$w_n(x,\tilde{y}) = \int_0^1 w(x,\tilde{y},z)\,m_n(z)\,dz,$$
$$g_{in}(x,\tilde{y}) = \int_0^1 g_i(x,\tilde{y},z)\,m_n(z)\,dz \quad \text{for } i = 1,2,$$

and define u_n, v_n, f_n, and g_{3n} as similar products with l_n. Multiply (3.17) and (3.20) by $l_n(z)$, (3.18) and (3.19) by $m_n(z)$, and integrate with respect to z. We obtain

$$u_{n,x} + v_{n,\tilde{y}} + n\pi\, w_n = f_n, \qquad n \ge 0, \tag{3.25}$$

$$w_{n,\tilde{y}} + n\pi\, v_n = g_{1n}, \qquad n \geq 1, \tag{3.26}$$

$$-\, n\pi\, u_n - w_{n,x} = g_{2n}, \qquad n \geq 1, \tag{3.27}$$

$$v_{n,x} - u_{n,\tilde{y}} = g_{3n}, \qquad n \geq 0. \tag{3.28}$$

Since $(f, \vec{g}, q, c)$ is in Z_0, $\vec{g}$ satisfies (3.16). Multiplying (3.16) by m_n and integrating with respect to z, we have

$$g_{1n,x} + g_{2n,\tilde{y}} - n\pi\, g_{3n} = 0 \quad \text{for } n \geq 1.$$

This implies, for $n \geq 1$, that (3.28) is automatically satisfied when (3.26) and (3.27) hold. Hence, for $n \geq 1$, u_n, v_n, and w_n can be determined by solving (3.25)–(3.27) by expanding in Fourier series in x and $\tilde{y}$. Similarly, we can determine u_0 and v_0 by solving (3.25), (3.28) with normalization (3.23). Now define

$$u(x, \tilde{y}, z) = \sum_{n=0}^{+\infty} u_n(x, \tilde{y})\, l_n(z),$$

$$v(x, \tilde{y}, z) = \sum_{n=0}^{+\infty} v_n(x, \tilde{y})\, l_n(z),$$

$$w(x, \tilde{y}, z) = \sum_{n=1}^{+\infty} w_n(x, \tilde{y})\, m_n(z).$$

It can be shown as in Lemma 5 in [1] that (u, v, w) thus defined solves (3.17)–(3.23) and satisfies estimate (3.24) with $q = 0$. Finally, it remains to remove the assumption $q = 0$. Given $(f, \vec{g}, q, c) \in Z_0$, we can find $\tilde{q} \in H^{s+1}_{o,e}(\tilde{\Omega})$ so that $\tilde{q} = q$ on $z = 1$, $\tilde{q} = 0$ on $z = 0$, and

$$|\, \tilde{q}\, |_{s+1} \leq C\, |\, q\, |_{s+1/2}.$$

Now if we set $w = \tilde{w} + \tilde{q}$, then the problem of solving for $(u, v, \tilde{w})$ brings us back to the previous case and we are done.

Since both J and L depend smoothly on η, θ, and ϵ, it follows from Lemma 3.2 that for η in a bounded set in $H^{s+1/2}_{e,e}(\tilde{\Gamma})$ and θ close to 1, $JL(\epsilon\eta, \theta)$ as an operator from Y to Z_0 is invertible for ϵ sufficiently small. Consequently, the solution of (3.12) is given by

$$(u, v, w) = (\, JL(\epsilon\,\eta, \theta)\,)^{-1} \left(0, \vec{0},\, \eta_x,\, -\int_{-\ell_2}^{\ell_2} \eta(0, \tilde{y})\, d\tilde{y}\right) \tag{3.29}$$

and is depending smoothly on η, θ, and ϵ.

Now because of (3.29), we are led to solve (3.8) as a nonlinear equation for η and θ. For ϵ close to zero, we seek η of the form

$$\eta = \cos k_1 x\, \cos k_2 \tilde{y} + \tilde{\eta} \tag{3.30}$$

in which $\tilde{\eta}$ is even in x and $\tilde{y}$, such that

$$\iint_{\widetilde{\Gamma}} \tilde{\eta}(x,\tilde{y}) \cos k_1 x \cos k_2 \tilde{y}\, dx d\tilde{y} = 0. \tag{3.31}$$

Let $X^{s+3/2}(\widetilde{\Gamma})$ be the subspace of $H^{s+3/2}_{e,e}(\widetilde{\Gamma})$ consistingof $\tilde{\eta}$ that satisfies (3.31). We can write Bernoulli's equation (3.8) as

$$B(\tilde{\eta}, \theta, \epsilon) = 0, \tag{3.32}$$

where B is a smooth function from an open subset of $X^{s+3/2}(\widetilde{\Gamma}) \times \mathbf{R}^2$ to $H^{s-1/2}_{e,e}(\widetilde{\Gamma})$. When $\epsilon = 0$, (3.32) has a trivial solution $\tilde{\eta} = 0$, $\theta = 1$. In what follows, we will assume that the wave numbers k_1 and k_2 satisfy the condition

$$b(mk_1, nk_2) \neq 0 \quad \text{for all } (m,n) \quad \text{except for } (m,n) = (1,1). \tag{3.33}$$

Here $b(\cdot,\cdot)$ is the function defined in (3.3). Based on (3.33), we will show that the Fréchet derivative

$$B_{(\tilde{\eta},\theta)}(0,1,0) : X^{s+3/2}(\widetilde{\Gamma}) \times \mathbf{R} \longrightarrow H^{s-1/2}_{e,e}(\widetilde{\Gamma})$$

has a bounded inverse.

In §4, we will limit the range of wave number k_1 to

$$\max\left(k_G, \frac{k_T}{2}\right) < k_1 < k_T.$$

When $\beta < 0.02$ and the Froude number is within the range $\gamma_0(\beta) < \gamma < 1$, Lemma 3.1 implies that equation $b(k_1, \xi) = 0$ has a unique positive root $\xi = k_2$ which is simple. For $m \neq 1$, mk_1 lies outside the range $[k_G, k_T]$ and so equation $b(mk_1, \xi) = 0$ has no real roots. As a result, condition (3.33) is met. For a detailed discussion of the set of excluded parameters, see §7 in [5].

When $\epsilon = 0$,

$$B(\tilde{\eta}, \theta, 0) = \eta - \beta \widetilde{\nabla}^2 \eta + \gamma u, \tag{3.34}$$

where $\widetilde{\nabla} = (\partial_x, \theta\, \partial_{\tilde{y}})$ and η is given by (3.30). (u,v,w) is a function of $\tilde{\eta}$ and θ, determined by

$$u_x + \theta\, v_{\tilde{y}} + w_z = 0, \tag{3.35}$$

$$\begin{aligned} \theta\, w_{\tilde{y}} - v_z &= 0, \\ u_z - w_x &= 0, \\ v_x - \theta\, u_{\tilde{y}} &= 0, \end{aligned} \tag{3.36}$$

$$w = 0 \quad \text{on } z = 0, \tag{3.37}$$

$$w = \eta_x \quad \text{on } z = 1, \tag{3.38}$$

$$\int_{-\ell_2}^{\ell_2} \int_0^1 u\, dz d\tilde{y} = -\int_{-\ell_2}^{\ell_2} \eta\, d\tilde{y} \quad \text{on } x = 0. \tag{3.39}$$

Given any $\tilde{\eta} \in X^{s+3/2}(\tilde{\Gamma})$, because of (3.31), we can express $\tilde{\eta}$ as

$$\tilde{\eta} = \sum_{(m,n)\neq(1,1)} a_{mn} \cos mk_1 x \cos nk_2 \tilde{y}.$$

As a result, η given by (3.30) has Fourier expansion

$$\eta = \sum_{m,n=0}^{+\infty} a_{mn} \cos mk_1 x \cos nk_2 \tilde{y},$$

where $a_{11} = 1$. By solving (3.35)–(3.39), we obtain

$$(3.40) \qquad u = -a_{00} - \sum_{(m,n)\neq(0,0)} a_{mn} \frac{m^2 {k_1}^2}{\psi_{mn}} \frac{\cosh \psi_{mn} z}{\sinh \psi_{mn}} \cos mk_1 x \cos nk_2 \tilde{y},$$

where $\psi_{mn} = (m^2 {k_1}^2 + \theta^2 n^2 {k_2}^2)^{1/2}$. It follows that

$$(3.41) \quad B(\tilde{\eta}, \theta, 0) = (1-\gamma)\, a_{00} - \sum_{(m,n)\neq(0,0)} b(mk_1, \theta nk_2)\, a_{mn} \cos mk_1 x \cos nk_2 \tilde{y}.$$

Note that the derivative $B_{(\tilde{\eta},\theta)}(0,1,0)$ is given by

$$(\tilde{\eta}, \tilde{\theta}) \longmapsto B(\tilde{\eta}, 1, 0) + \tilde{\theta}\, B_\theta(0,1,0).$$

When $\theta = 1$, the coefficient $b(k_1, k_2)$ in (3.41) is equal to zero. This implies that $B(\cdot, 1, 0)$ is a linear operator from $X^{s+3/2}(\tilde{\Gamma})$ to $X^{s-1/2}(\tilde{\Gamma})$. From (3.3), it is easy to see that, when $\mid \tau \mid$ is sufficiently large, coth $\mid \tau \mid$ is close to 1 and the term $\beta \mid \tau \mid^2$ becomes dominant in $b(\tau_1, \tau_2)$. Thus, for m and n sufficiently large, we have

$$\mid b(mk_1, nk_2) \mid \geq C\, (m^2 {k_1}^2 + n^2 {k_2}^2).$$

As a result, $B(\cdot, 1, 0)$ has a bounded inverse from $X^{s-1/2}(\tilde{\Gamma})$ to $X^{s+3/2}(\tilde{\Gamma})$.

Next we compute the derivative $B_\theta(0,1,0)$. When $\tilde{\eta} = 0$, (3.40) gives

$$u = -\frac{{k_1}^2}{\psi_{11}} \coth \psi_{11} \cos k_1 x \cos k_2 \tilde{y} \quad \text{on } z = 1,$$

where $\psi_{11} = ({k_1}^2 + \theta^2 {k_2}^2)^{1/2}$. Note that, given a smooth function f, we have

$$\{ \partial_\theta f(\theta k_2) \}_{\theta=1} = \{ \xi \partial_\xi f(\xi) \}_{\xi=k_2}.$$

This simple fact leads to

$$u_\theta = \left\{ \xi \partial_\xi \left(\frac{-{k_1}^2 \coth \sqrt{{k_1}^2 + \xi^2}}{\sqrt{{k_1}^2 + \xi^2}} \right) \right\}_{\xi = k_2} \cos k_1 x \cos k_2 \tilde{y}.$$

Hence, from (3.34), we obtain

$$\begin{aligned} B_\theta(0,1,0) &= 2\beta {k_2}^2 \cos k_1 x \cos k_2 \tilde{y} + \gamma u_\theta \\ &= k_2 \{ \partial_\xi b(k_1, \xi) \}_{\xi = k_2} \cos k_1 x \cos k_2 \tilde{y}. \end{aligned}$$

From Lemma 3.1, since k_2 is a simple root of $b(k_1, \xi) = 0$, $\{\partial_\xi\, b(k_1, \xi)\}_{\xi=k_2} \neq 0$. Thus the derivative $B_\theta(0, 1, 0)$ is nonzero.

Now let $f \in H^{s-1/2}_{e,e}(\widetilde{\Gamma})$ be arbitrary and consider the equation

$$B(\tilde{\eta}, 1, 0) + \tilde{\theta}\, B_\theta(0, 1, 0) = f.$$

Since the operator $B(\cdot, 1, 0)$ ranges in $X^{s-1/2}(\widetilde{\Gamma})$ and $B_\theta(0, 1, 0)$ is orthogonal to $X^{s-1/2}(\widetilde{\Gamma})$ in the L^2 sense, $\tilde{\theta}$ should be chosen so that

$$f - \tilde{\theta}\, B_\theta(0, 1, 0) \in X^{s-1/2}(\widetilde{\Gamma}).$$

Then

$$B(\tilde{\eta}, 1, 0) = f - \tilde{\theta}\, B_\theta(0, 1, 0)$$

has a unique solution $\tilde{\eta}$ in $X^{s+3/2}(\widetilde{\Gamma})$. It is easy to see that the Fréchet derivative $B_{(\tilde{\eta},\theta)}(0, 1, 0)$ is bijective. Thus, by the open mapping theorem, $B_{(\tilde{\eta},\theta)}(0, 1, 0)$ has a bounded inverse.

Now by the implicit function theorem, for each ϵ sufficiently close to zero, (3.32) has a solution $(\tilde{\eta}(\epsilon), \theta(\epsilon))$ near $(0, 1)$ in $X^{s+3/2}(\widetilde{\Gamma}) \times \mathbf{R}$, depending smoothly on ϵ. By (3.29) and (3.30), $(\tilde{\eta}(\epsilon), \theta(\epsilon))$ thus obtained provides us with an exact doubly periodic solution of (3.4)–(3.8) and (3.11).

Finally, we point out that given any constant a in a bounded set, if η, θ, and (u, v, w) satisfy (3.4)–(3.8) with ϵ replaced by $a\epsilon$, then $a\eta(a\epsilon)$, $\theta(a\epsilon)$, and $au(a\epsilon)$, $av(a\epsilon)$, $aw(a\epsilon)$ also satisfy the same problem. Moreover, if we translate the $(x, \tilde{y})$ coordinates by $(0, \delta)$, then

$$\begin{aligned} &a\eta(x, \tilde{y} - \delta; a\epsilon), \quad \theta(a\epsilon), \quad au(x, \tilde{y} - \delta, z; a\epsilon), \\ &av(x, \tilde{y} - \delta, z; a\epsilon), \qquad aw(x, \tilde{y} - \delta, z; a\epsilon) \end{aligned} \tag{3.42}$$

still satisfy (3.4)–(3.8) with ϵ replaced by $a\epsilon$. By construction, when $\epsilon = 0$, we have

$$a\eta(x, \tilde{y} - \delta; 0) = a \cos k_1 x \cos k_2(\tilde{y} - \delta). \tag{3.43}$$

Thus, by treating the constants a and δ as two free parameters, we obtain a two-parameter family of doubly periodic solutions. The constants a and δ can be interpreted as amplitude and phase shift, respectively.

4. Flows due to a partially localized pressure disturbance.

We now turn to the main issue of this paper. Suppose that, in (2.12), we are given a pressure disturbance p which is periodic in x with period $2\ell_1 = 2\pi / k_1$ and is in ${}_\rho H^{s-1/2}_e(\Gamma)$ for some integer $s > \frac{3}{2}$. The problem is to solve (2.8)–(2.12) for η and (u, v, w), provided ϵ is sufficiently small. Note that, by definition of the ${}_\rho H^s$ spaces, the pressure disturbance p is partially localized in the sense that p and its derivatives are decaying rapidly as $y \longrightarrow \pm\infty$. In the following discussion, we will assume that the wave number k_1 is in the range

$$\max\left(k_G, \frac{k_T}{2}\right) < k_1 < k_T. \tag{4.1}$$

Under this circumstance, when the surface tension effect is small and the flow speed U_0 is within a certain range, we are able to use the symmetric doubly periodic waves discussed in §3 to construct solutions of (2.8)–(2.12).

4.1. A linearized problem. We first construct linear approximations by solving (2.8)–(2.12) with $\epsilon = 0$. Consider

$$u_x + v_y + w_z = 0, \tag{4.2}$$

$$\begin{aligned} w_y - v_z &= 0, \\ u_z - w_x &= 0, \\ v_x - u_y &= 0, \end{aligned} \tag{4.3}$$

$$w = 0 \quad \text{on } z = 0, \tag{4.4}$$

$$w = \eta_x \quad \text{on } z = 1, \tag{4.5}$$

$$\eta - \beta \nabla^2 \eta + \gamma u + p = 0 \quad \text{on } z = 1, \tag{4.6}$$

where $\nabla = (\partial_x, \partial_y)$. Assume that, in (4.6), the given pressure disturbance p has the form

$$p = \sum_{m=0}^{+\infty} p_m(y) \cos mk_1 x.$$

We seek a solution of (4.2)–(4.6) which has η even in x. Let

$$\eta = \sum_{m=0}^{+\infty} \eta_m(y) \cos mk_1 x.$$

By applying Fourier series in the x direction and Fourier transform in the y direction, we can solve (4.2)–(4.5) and regard (u, v, w) as a function of η. We obtain

$$\widehat{u}(x, \xi, z) = \sum_{m=1}^{+\infty} -\frac{m^2 {k_1}^2}{\psi_m(\xi)} \widehat{\eta_m}(\xi) \frac{\cosh \psi_m(\xi) z}{\sinh \psi_m(\xi)} \cos mk_1 x,$$

where $\psi_m(\xi) = (m^2 {k_1}^2 + \xi^2)^{1/2}$, $\widehat{}$ is the Fourier transform in y with dual variable ξ. It remains to solve (4.6) as a linear equation for η. From (4.6), by taking Fourier transform with respect to y, we see that η solves (4.6) if

$$\sum_{m=0}^{+\infty} \{b(mk_1, \xi)\, \widehat{\eta_m}(\xi) + \widehat{p_m}(\xi)\} \cos mk_1 x = 0.$$

Here $b(\cdot, \cdot)$ is the function given in (3.3). Hence we are led to solve

$$b(mk_1, \xi)\, \widehat{\eta_m}(\xi) + \widehat{p_m}(\xi) = 0 \tag{4.7}$$

for each integer $m \geq 0$.

We now discuss properties of the function $b(\cdot, \cdot)$ in more detail.

LEMMA 4.1. *Let ρ be any positive constant $< k_1$. For each integer $m \geq 0$, the function*

$$b(mk_1, z) = 1 + \beta\, (m^2 {k_1}^2 + z^2) - \gamma \frac{m^2 {k_1}^2}{\sqrt{m^2 {k_1}^2 + z^2}} \coth \sqrt{m^2 {k_1}^2 + z^2}$$

is an analytic function in z over the strip $\{z : |\operatorname{Im} z| \leq \rho\}$. Furthermore,

$$|b(mk_1, z)| \geq c\,(m^2{k_1}^2 + |\operatorname{Re} z|^2)$$

when m or $\operatorname{Re} z$ is sufficiently large. Here c is a positive constant independent of m.

Note that, for any complex number ω,

$$|\coth \omega|^2 = \frac{\sinh^2 \operatorname{Re}\omega + \cos^2 \operatorname{Im}\omega}{\sinh^2 \operatorname{Re}\omega + \sin^2 \operatorname{Im}\omega}.$$

It is clear that $|\coth\sqrt{m^2{k_1}^2 + z^2}|$ is close to 1 provided m or $\operatorname{Re} z$ is sufficiently large. Based on this observation, the proof of Lemma 4.1 is straightforward and is thus omitted. The following corollary concerns roots of analytic functions $b(mk_1, z)$.

COROLLARY 4.2. *Suppose parameters β and γ are the same as in Lemma 3.1. Given $k_1 > 0$ such that*

$$\max\left(k_G, \frac{k_T}{2}\right) < k_1 < k_T,$$

there exists a positive number $\rho_* < k_1$ such that

(i) *for each nonnegative integer $m \neq 1$, $b(mk_1, z)$ is nonzero for all z in the strip $|\operatorname{Im} z| \leq \rho_*$;*

(ii) *$b(k_1, z)$ has two roots in the strip $|\operatorname{Im} z| \leq \rho_*$, all real and simple.*

Proof. See Appendix. □

The following lemma is used to solve (4.7).

LEMMA 4.3. *Suppose b is an analytic function from a strip $\{\xi + i\phi : |\phi| \leq \rho\}$ to $\mathbf{C}$ satisfying $|b(\xi + i\phi)| \geq c_0 |\xi|^m$ for ξ sufficiently large, where $m \geq 0$, $\rho > 0$. Suppose that b has only a finite number of roots, all real and simple. Given $f \in {}_\rho H^s(\mathbf{R})$, $s \geq 0$ with the property that $\hat{f}(\xi) = 0$ at each ξ for which $b(\xi) = 0$, we define η as the inverse transform of $\hat{f}(\xi)/b(\xi)$. Then η is the unique function in $L^1(\mathbf{R})$ satisfying $b\hat{\eta} = \hat{f}$; $\eta \in {}_\rho H^{s+m}(\mathbf{R})$ and*

$$|\eta|_{s+m,\rho} \leq C|f|_{s,\rho}.$$

Here the constant C depends on b but not on f.

Proof. See Lemma 3 in [1]. □

In what follows, beside (4.1), we will also assume that $p \in {}_\rho H_e^{s-1/2}(\Gamma)$ where $\rho < \rho_*$. Here ρ_* is the constant that appears in Corollary 4.2. Under these assumptions, with $0 < \beta < 0.02$ and $\gamma_0(\beta) < \gamma < 1$, we are able to show that (4.2)–(4.6) have a two-parameter family of solutions which exhibit doubly periodic wave patterns at infinity. See (4.17), (4.18) below.

Since $p \in {}_\rho H_e^{s-1/2}(\Gamma)$, in (4.7), $p_m(y)$ is in ${}_\rho H^{s-1/2}(\mathbf{R})$ for each $m \geq 0$. Now, for each nonnegative integer $m \neq 1$, Corollary 4.2 implies that $b(mk_1, z) \neq 0$ for all z in the strip $|\operatorname{Im} z| \leq \rho$. Thus, by Lemmas 4.1 and 4.3, there exists a $\eta_m(y)$ in ${}_\rho H^{s+3/2}(\mathbf{R})$ that satisfies (4.7). As for $m = 1$, Corollary 4.2 shows that $b(k_1, z) = 0$ has only two roots $\xi = \pm k_2$, all real and simple. We will show that, when $m = 1$, (4.7) has a solution η_1 of the form

$$\eta_1 = \eta_1^+ + \eta_1^0 + \eta_1^- \tag{4.8}$$

in which $\eta_1^\pm(y) = a^\pm \zeta^\pm(y) \cos k_2(y - \delta^\pm)$ and $\eta_1^0 \in {}_\rho H^{s+3/2}(\mathbf{R})$. Here $\zeta^\pm(y)$ are two C^∞ cut-off functions such that $0 \leq \zeta^+ \leq 1$, $\zeta^+(y) = 1$ for $y \geq 1$, $\zeta^+(y) = 0$ for $y \leq -1$, and $\zeta^-(y) = 1 - \zeta^+(y)$.

From distribution theory, we know that the Fourier transforms of $\zeta^{\pm}$ are

$$\widehat{\zeta^{\pm}}(\xi) = \sqrt{\frac{\pi}{2}}\,\delta(\xi) \pm \frac{\widehat{\chi}(\xi)}{i\,\xi}, \tag{4.9}$$

in which $\delta(\xi)$ is the delta function and $\chi = \partial_y\,\zeta^+$. It is immediate from (4.9) that the transforms of $\eta_1^{\pm}$ are

$$\begin{aligned} \widehat{\eta_1^{\pm}}(\xi) = {} & \pm \frac{a^{\pm}\, e^{i\,k_2\,\delta^{\pm}}}{2i\,(\xi + k_2)}\,\widehat{\chi}(\xi + k_2) \pm \frac{a^{\pm}\, e^{-i\,k_2\,\delta^{\pm}}}{2i\,(\xi - k_2)}\,\widehat{\chi}(\xi - k_2) \\ & + \sqrt{\frac{\pi}{8}}\left[e^{i\,k_2\,\delta^{\pm}}\,\delta(\xi + k_2) + e^{-i\,k_2\,\delta^{\pm}}\,\delta(\xi - k_2)\right]. \end{aligned} \tag{4.10}$$

From (4.10), since $\pm k_2$ are simple roots of $b(k_1, \xi) = 0$, we have

$$\begin{aligned} \left[b(k_1,\cdot)\,\widehat{\left(\eta_1^{\pm}\right)}\right](k_2) &= \pm a^{\pm}\, e^{-i\,k_2\,\delta^{\pm}}\; c, \\ \left[b(k_1,\cdot)\,\widehat{\left(\eta_1^{\pm}\right)}\right](-k_2) &= \mp a^{\pm}\, e^{i\,k_2\,\delta^{\pm}}\; c, \end{aligned} \tag{4.11}$$

where c is a nonzero, pure imaginary, complex number.

Now if a solution of the form (4.8) does exist, then we have

$$b(k_1,\cdot)\,\widehat{\left(\eta_1^{0}\right)} = -\widehat{p_1} - b(k_1,\cdot)\,\widehat{\left(\eta_1^{+}\right)} - b(k_1,\cdot)\,\widehat{\left(\eta_1^{-}\right)}. \tag{4.12}$$

Since η_1^0 decays rapidly as $y \longrightarrow \pm\infty$, its transform $\widehat{\eta_1^0}$ is a smooth function of ξ and so

$$\left[b(k_1,\cdot)\,\widehat{\left(\eta_1^{0}\right)}\right](\pm\,k_2) = 0.$$

Hence we are led to choose the amplitudes $a^{\pm}$ such that

$$\left[b(k_1,\cdot)\,\widehat{\left(\eta_1^{+}\right)}\right](k_2) + \left[b(k_1,\cdot)\,\widehat{\left(\eta_1^{-}\right)}\right](k_2) = -\widehat{p_1}(k_2),$$

that is to say,

$$a^{+}\, e^{-i\,k_2\,\delta^{+}} - a^{-}\, e^{-i\,k_2\,\delta^{-}} = -\frac{\widehat{p_1}(k_2)}{c} \tag{4.13}$$

according to (4.11). Here c is a nonzero, pure imaginary, complex number. The real and imaginary parts of the complex numbers in (4.13) form a linear system of two equations in $a^{\pm}$, whose determinant is $-\sin k_2(\delta^+ - \delta^-)$. Given phase shifts $\delta^{\pm}$ so that $\sin k_2(\delta^+ - \delta^-) \neq 0$, we can uniquely determine the solution $a_0^{\pm}$ of (4.13). Clearly, the amplitudes $a_0^{\pm}$ satisfy

$$\mid a_0^{\pm} \mid \leq C \mid \widehat{p_1}(k_2) \mid \tag{4.14}$$

for some constant C. Note that, since $p_1(y)$ is real-valued, $\widehat{p_1}(-k_2)$ is equal to the complex conjugate of $\widehat{p_1}(k_2)$. From (4.11), it is easy to see that the amplitudes $a_0^{\pm}$ determined above also satisfy

$$\left[b(k_1,\cdot)\,\widehat{\left(\eta_1^{+}\right)}\right](-k_2) + \left[b(k_1,\cdot)\,\widehat{\left(\eta_1^{-}\right)}\right](-k_2) = -\widehat{p_1}(-k_2).$$

Consequently, the right-hand side of (4.12) vanishes at $\xi = \pm k_2$.

Now (4.12) can be rewritten as

$$b(k_1,\cdot)\,\widehat{(\eta_1^0)} = \widehat{f}, \tag{4.15}$$

where

$$f = -p_1 - \left[b(k_1,\cdot)\,(\eta_1^+)^\wedge\right]^\vee - \left[b(k_1,\cdot)\,(\eta_1^-)^\wedge\right]^\vee.$$

Note that, by construction, $\widehat{f}(\pm k_2) = 0$. Moreover $b(k_1,\xi)\,(\eta_1^\pm)^\wedge(\xi)\cos k_1x$ are the transforms of the left-hand side of (4.6) with $p = 0$ and η replaced by $\eta_1^\pm(y)\cos k_1x$. From §3, we know that $a_0^\pm \cos k_1x \cos k_2(y - \delta^\pm)$ are the surface elevations of two linear symmetric doubly periodic waves. Consequently, $\eta_1^\pm(y)\cos k_1x$ satisfies (4.6) with $p = 0$ for $|y| \geq 1$, and we can regard $[b(k_1,\cdot)\,(\eta_1^\pm)^\wedge]^\vee$ as two smooth functions of y with compact support in $|\,y\,| \leq 1$. Thus the function f defined above is in ${}_\rho H^{s-1/2}(\mathbf{R})$ with

$$|\,f\,|_{s-1/2,\rho} \leq C(\,|\,p_1\,|_{s-1/2,\rho} + |\,\widehat{p_1}(k_2)\,|\,).$$

By Lemma 4.3, (4.15) has a solution η_1^0 in${}_\rho H^{s+3/2}(\mathbf{R})$ and

$$|\,\eta_1^0\,|_{s+3/2,\rho} \leq C(\,|\,p_1\,|_{s-1/2,\rho} + |\,\widehat{p_1}(k_2)\,|\,). \tag{4.16}$$

With all the η_m determined, for each pair of phase shifts $\delta^\pm$ such that

$$\sin k_2(\delta^+ - \delta^-) \neq 0,$$

we obtain a solution η of the form (4.6) of the form

$$\begin{aligned}\eta = {} & a_0^+\,\zeta^+(y)\cos k_1x\,\cos k_2(y - \delta^+) + \eta^* \\ & + a_0^-\,\zeta^-(y)\cos k_1x\,\cos k_2(y - \delta^-),\end{aligned} \tag{4.17}$$

where

$$\eta^* = \eta_1^0(y)\cos k_1x + \sum_{m\neq 1} \eta_m(y)\cos mk_1x. \tag{4.18}$$

To complete our discussion of the linearized problem, we will show that η^* is in ${}_\rho H_e^{s+3/2}(\Gamma)$.

To estimate $|\,\eta^*\,|_{s+3/2,\rho}$, it suffices to consider $|e^{\pm\rho y}\eta^*|_{s+3/2}$. Note that

(4.19)

$$\begin{aligned}|\,e^{\pm\rho y}\eta^*\,|_{s+3/2} = {} & \int_{-\infty}^{+\infty} (1 + k_1{}^2 + \xi^2)^{s+3/2}|(e^{\pm\rho y}\eta_1^0)^\wedge(\xi)|^2\,d\xi \\ & + \sum_{m\neq 1}\int_{-\infty}^{+\infty} (1 + m^2k_1{}^2 + \xi^2)^{s+3/2}\,|\,(e^{\pm\rho y}\eta_m)^\wedge(\xi)\,|^2\,d\xi.\end{aligned}$$

Here the first term in (4.19) is finite because of (4.16). As in the proof of the Paley–Wiener theorem, given a function $f \in {}_\rho H^s(\mathbf{R})$, its Fourier transform $\widehat{f}$ can be extended to an analytic function over the strip $\{\xi + i\phi : |\,\phi\,| < \rho\}$ by defining

$$\widehat{f}(\xi + i\phi) = \widehat{(e^{\phi x} f)}(\xi).$$

Moreover, $\widehat{f}$ has boundary values in the sense $\widehat{f}(\cdot + i\phi) \to \widehat{f}(\cdot \pm i\rho)$ in $L^2(\mathbf{R})$ as $\phi \to \pm\rho$. See [6]. For each nonnegative integer $m \neq 1$, we can extend $\eta_m(y)$ and $p_m(y)$ into analytic functions over the strip $|\operatorname{Im} z| \leq \rho$, where $\rho < \rho_*$. By Corollary 4.2, for each such m, $b(mk_1, z) \neq 0$ for any z in $|\operatorname{Im} z| \leq \rho$. Now (4.7) implies that $\widehat{p_m}(z) / b(mk_1, z)$ is an analytic function over the strip and is equal to the $\widehat{\eta_m}(z)$ along the real line. Thus, for each nonnegative integer $m \neq 1$, we have

$$\widehat{\eta_m}(z) = \widehat{p_m}(z) / b(mk_1, z)$$

for all z in the strip $|\operatorname{Im} z| \leq \rho$. Now, by Lemma 4.1, the second term in (4.19) can be estimated as follows:

$$\begin{aligned}
&\sum_{m\neq 1} \int_{-\infty}^{+\infty} (1 + m^2k_1{}^2 + \xi^2)^{s+3/2} \,|\, (e^{\pm \rho y}\eta_m)\widehat{\ }(\xi) \,|^2 \, d\xi \\
&= \sum_{m\neq 1} \int_{-\infty}^{+\infty} (1 + m^2k_1{}^2 + \xi^2)^{s+3/2} \,|\, \widehat{\eta_m}(\xi \pm i\rho) \,|^2 \, d\xi \\
&= \sum_{m\neq 1} \int_{-\infty}^{+\infty} (1 + m^2k_1{}^2 + \xi^2)^{s+3/2} \frac{|\, \widehat{p_m}(\xi \pm i\rho) \,|^2}{|\, b(mk_1, \, \xi \pm i\rho) \,|^2} \, d\xi \\
&\leq C \sum_{m\neq 1} \int_{-\infty}^{+\infty} (1 + m^2k_1{}^2 + \xi^2)^{s-1/2} \,|\, \widehat{p_m}(\xi \pm i\rho) \,|^2 \, d\xi \\
&\leq C \,|\, p \,|_{s-1/2,\rho}^2.
\end{aligned}$$

The above estimate and (4.16) show that η^* is in ${}_\rho H_e^{s+3/2}(\Gamma)$, where

$$|\, \eta^* \,|_{s+3/2,\rho} \leq C(\,|\, p \,|_{s-1/2,\rho} + |\, \widehat{p_1}(k_2) \,|\,). \tag{4.20}$$

Since $p_1(y)$ is in ${}_\rho H^{s-1/2}(\mathbf{R})$, where $s > \frac{3}{2}$, $|\, \widehat{p_1}(k_2) \,|$ is then bounded above by $|p_1|_{s-1/2,\rho}$. Thus, from (4.14) and (4.20), we obtain

$$|\, \eta^* \,|_{s+3/2,\rho} + |\, a_0^+ \,| + |\, a_0^- \,| \leq C \,|\, p \,|_{s-1/2,\rho} \tag{4.21}$$

for some constant C.

4.2. The nonlinear problem. We now show that we can use the exact symmetric doubly periodic waves of §3 to construct exact solutions of (2.8)–(2.12), which exhibit doubly periodic wave patterns as $y \to \pm\infty$. Note that the symmetric doubly periodic waves considered have wave number $\theta(\epsilon)\, k_2$ in the y direction. For the same reason as in §3, we stretch the y coordinate by setting $\tilde{y} = \theta(\epsilon)\, y$. Here $\theta(\epsilon)$ is the stretching factor determined in §3. After the stretching, we are led to consider

$$\{(1 + \epsilon\eta)\, u_x - \epsilon\, z\, \eta_x\, u_z\} + \theta\, \{(1 + \epsilon\eta)\, v_{\tilde{y}} - \epsilon\, z\, \eta_{\tilde{y}}\, v_z\} + w_z = 0, \tag{4.22}$$

$$\begin{aligned}
\theta\, w_{\tilde{y}} - \sigma\, v_z - \epsilon\, \theta\, z\, \sigma\, \eta_{\tilde{y}}\, w_z &= 0, \\
\sigma\, u_z - w_x + \epsilon\, z\, \sigma\, \eta_x\, w_z &= 0, \\
v_x - \theta\, u_{\tilde{y}} - \epsilon\, z\, \sigma\, \eta_x\, v_z + \epsilon\, \theta\, z\, \sigma\, \eta_{\tilde{y}}\, u_z &= 0,
\end{aligned} \tag{4.23}$$

$$w - \epsilon\, u\, \eta_x - \epsilon\, \theta\, v\, \eta_{\tilde{y}} = \eta_x \quad \text{on } z = 1, \tag{4.24}$$

(4.25)
$$w = 0 \quad \text{on } z = 0,$$

(4.26)
$$\eta - \beta \widetilde{\nabla} \cdot \frac{\widetilde{\nabla} \eta}{[1 + \epsilon^2 \mid \widetilde{\nabla} \eta \mid^2]^{1/2}} + \gamma u + \frac{\epsilon}{2} \gamma (u^2 + v^2 + w^2) + p = 0 \quad \text{on } z = 1.$$

Here $\sigma = (1 + \epsilon \eta)^{-1}$ and $\widetilde{\nabla} = (\partial_x, \theta \partial_{\tilde{y}})$.

We seek a surface elevation η of the form

(4.27)
$$\eta = \eta^+ + \eta^0 + \eta^-$$

in which $\eta^0 \in {}_\rho H_e^{s+3/2}(\Gamma)$ and

$$\eta^\pm = \zeta^\pm(\tilde{y})\, \eta_*^\pm(x, \tilde{y}, a^\pm, \delta^\pm, \epsilon).$$

Here $\eta_*^\pm$ are the doubly periodic surface elevations in (3.42). $\zeta^\pm(\tilde{y})$ are C^∞ cut-off functions so that $0 \leq \zeta^+ \leq 1$, $\zeta^+(\tilde{y}) = 1$ for $\tilde{y} > 1$, $\zeta^+(\tilde{y}) = 0$ for $\tilde{y} < -1$, and $\zeta^-(\tilde{y}) = 1 - \zeta^+(\tilde{y})$. Similarly, we seek (u, v, w) of the form

(4.28)
$$\begin{aligned} u &= u^+ + u^0 + u^-, \\ v &= v^+ + v^0 + v^-, \\ w &= w^+ + w^0 + w^-, \end{aligned}$$

in which

$$(u_0, v_0, w_0) \in {}_\rho H_e^{s+1}(\Omega) \times [{}_\rho H_o^{s+1}(\Omega)]^2$$

and

$$(u^\pm, v^\pm, w^\pm) = (\zeta^\pm u_*^\pm, \zeta^\pm v_*^\pm, \zeta^\pm w_*^\pm).$$

Here $(u_*^\pm, v_*^\pm, w_*^\pm)$ are the correction terms of the doubly periodic velocities in (3.42), corresponding to $\eta_*^\pm$.

We now summarize the main result of this paper. Suppose that $0 < \beta < 0.02$. In case of air-water interface at 20°C, with standard value for g, this corresponds to $h >$ 10cm. Furthermore, assume that the Froude number γ is within the range $\gamma_0(\beta) < \gamma < 1$ and a pressure disturbance $p \in {}_\rho H_e^{s-1/2}(\Gamma)$ with $s > \frac{3}{2}$ is given. Let

$$p(x, \tilde{y}) = \sum_{m=0}^{+\infty} p_m(\tilde{y}) \cos m k_1 x.$$

We consider the case in which p's wave number k_1 in the x direction satisfies

$$\max(k_G, k_T/2) < k_1 < k_T$$

and $0 < \rho < \rho_*$. The constant ρ_* is described in Corollary 4.2. We have the following.

THEOREM 4.4. *With the assumptions above, let k_2 be the positive root of equation $b(k_1, \xi) = 0$. When the Fourier transform $\widehat{p_1}$is nonzero at k_2, for each pair of phase shifts $\delta^\pm$ such that* $\sin k_2(\delta^+ - \delta^-) \neq 0$, (4.22)–(4.26) *has a solution $\eta(\epsilon)$,* $(u(\epsilon), v(\epsilon), w(\epsilon))$ *of the form* (4.27) *and* (4.28) *for each ϵ sufficiently small.*

From each solution of (4.22)–(4.26), we can define (U, V, W) and S according to (2.7). After returning to the original (X, Y, Z) coordinates, we obtain a two-parameter family of exact solutions of (2.1)–(2.5) which exhibit doubly periodic wave patterns as $Y \to \pm\infty$.

By multiplying all three equations in (4.23) by $(1 + \epsilon\,\eta)$, we can rewrite (4.22)–(4.24) as

$$\{L_0 + \epsilon\, L_1(\eta)\}(u, v, w) = (0, \vec{0}, \eta_x), \tag{4.29}$$

in which L_0 and $L_1(\eta)$ are linear operators, with $L_1(\eta)$ depending linearly on η. Suppose that (4.22)–(4.24) do have a solution η and (u, v, w) of the form (4.27) and (4.28). Note that we can rearrange (4.29) as

$$\begin{aligned}\{\, L_0 + \epsilon\, L_1(\eta)\,\}(u^0, v^0, w^0) &= (0, \vec{0}, \eta_x) - \{L_0 + \epsilon\, L_1(\eta)\}(u^\pm, v^\pm, w^\pm)\\ &= \Big[(0, \vec{0}, {\eta^\pm}_x) - \{L_0 + \epsilon\, L_1(\eta^\pm)\}(u^\pm, v^\pm, w^\pm)\Big]\\ &\quad - \epsilon\, L_1(\eta^\pm)(u^\mp, v^\mp, w^\mp) - \epsilon\, L_1(\eta^0)(u^\pm, v^\pm, w^\pm)\\ &\quad + (0, \vec{0}, {\eta^0}_x).\end{aligned} \tag{4.30}$$

Here each term with ± represents a sum of two terms. In (4.30), the first two terms on the right vanish for $|\ \tilde{y}\ | \geq 1$ since $\eta_*^\pm$ and $(u_*^\pm, v_*^\pm, w_*^\pm)$ satisfy (3.4), (3.5), and (3.7). The third and fourth terms decay rapidly as $\tilde{y} \to \pm\infty$ since $\eta^0 \in {}_\rho H_e^{s+3/2}(\Omega)$. Therefore, if we can invert $L_0{+}\epsilon\, L_1(\eta)$ over an appropriate space of functions which decay rapidly as $\tilde{y} \to \pm\infty$, then we can regard (u^0, v^0, w^0) as a function of η^0, $a^\pm$, $\delta^\pm$, and ϵ.

Let

$$\begin{aligned} Y &= \{(u, v, w) \in {}_\rho H_e^{s+1}(\Omega) \times [\,{}_\rho H_o^{s+1}(\Omega)\,]^2 \,:\, w(x, \tilde{y}, 0) = 0\},\\ Z &= [\,{}_\rho H_o^s(\Omega)\,]^2 \times [\,{}_\rho H_e^s(\Omega)\,]^2 \times {}_\rho H_o^{s+1/2}(\Gamma)\end{aligned}$$

and, for each η of the form (4.27), regard $L_0 + \epsilon\, L_1(\eta)$ as a linear operator from Y to Z. To take a closer look at the range of $L_0 + \epsilon\, L_1(\eta)$, suppose we have

$$\{\, L_0 + \epsilon\, L_1(\eta)\,\}(u, v, w) = (f, \vec{g}, q) \tag{4.31}$$

for some $(u, v, w) \in Y$. Here $(f, \vec{g}, q)$ is in Z with $\vec{g} = (g_1, g_2, g_3)$. As in (3.13), (4.31) implies that $\vec{g}$ is a curl in the original (X, Y, Z) coordinates. Therefore, (3.14) still holds for $\vec{g}$. Now if we define $\widetilde{g} = (\tilde{g}_1, \tilde{g}_2, \tilde{g}_3)$ as we did in (3.15), then $\widetilde{g}$ satisfies

$$\tilde{g}_{1,x} + \tilde{g}_{2,\tilde{y}} + \tilde{g}_{3,z} = 0. \tag{4.32}$$

Furthermore, by(3.15) and (4.31),

$$\begin{aligned}\tilde{g}_3 &= v_x - \theta\, u_{\tilde{y}} + \epsilon\,\theta\, z\,(\eta_{\tilde{y}}\, w_x - \eta_x\, w_{\tilde{y}}),\\ &= v_x - \theta\, u_{\tilde{y}} + \epsilon\,\theta\, z\,\{(\eta_{\tilde{y}}\, w)_x - (\eta_x\, w)_{\tilde{y}}\},\end{aligned} \tag{4.33}$$

which shows that $\tilde{g}_3$ is a divergence with respect to x, $\tilde{y}$. Motivated by (4.32) and (4.33), we define a subspace Z_0 of Z as follows. Let φ be a C^∞ function of compact support on $\mathbf{R}$ with $\varphi(0) = 1$. Let Z_0 be the subspace of Z consisting of $(f, \vec{g}, q)$ in which $\vec{g} = (g_1, g_2, g_3)$ satisfies

$$g_{1,x} + g_{2,\tilde{y}} + g_{3,z} = 0, \tag{4.34}$$

$$\int_{-\ell_1}^{+\ell_1} \widehat{g_3}(x,0,z)\,dx = 0, \tag{4.35}$$

$$\int_{-\infty}^{+\infty}\int_0^1 \frac{\varphi^2(\xi)}{\xi^2}\left\{\int_{-\ell_1}^{+\ell_1} \widehat{g_3}(x,\xi,z)\,dx\right\}^2 dzd\xi < +\infty. \tag{4.36}$$

Here $\widehat{\ }$ is the Fourier transform with respect to $\tilde{y}$ with dual variable ξ, and $\ell_1 = \pi/k_1$. Condition (4.36) states that the quantity in the brackets is zero at $\xi = 0$ in a generalized sense. The square of the norm in Z_0 is a set equal to the square of the norm in Z plus the term in (4.36).

As in §3, let $J(\epsilon\eta) : Z \to Z$ be the linear operator in (3.15) that maps $(f,\vec{g},q)$ to $(f,\widetilde{g},q)$. Now we can treat $J(\epsilon\eta)\,\{L_0 + \epsilon\, L_1(\eta)\}$ as a bounded operator from Y to Z_0. Clearly, when $\epsilon = 0$, the stretching factor $\theta = 1$ and thus $J(0)$ is the identity operator. In the following lemma, we show that $J(0)\,L_0 = L_0$ is invertible.

LEMMA 4.5. *Given any $(f,\vec{g},q)$ in Z_0, there exists a unique solution (u,v,w) in Y that satisfies*

$$u_x + v_{\tilde{y}} + w_z = f, \tag{4.37}$$

$$w_{\tilde{y}} - v_z = g_1, \tag{4.38}$$
$$u_z - w_x = g_2, \tag{4.39}$$
$$v_x - u_{\tilde{y}} = g_3, \tag{4.40}$$

$$w = q \quad on\ z = 1, \tag{4.41}$$

$$w = 0 \quad on\ z = 0, \tag{4.42}$$

and the estimate

$$|\,(u,v,w)\,|_Y \le C\,|\,(f,\vec{g},q)\,|_{Z_0}.$$

Proof. We use the method of Lemma 3.2 and start with the case $q = 0$. With u_n, v_n, w_n, f_n, and g_{in} defined as before, we are led to solve (3.25)–(3.28). According to their symmetries in the x direction, we can expand all functions involved into Fourier series in x. For example, we have

$$u_n = \sum_{m=0}^{+\infty} u_{nm}(\tilde{y})\,\cos mk_1x.$$

Similar to (3.16), (4.34) implies

$$g_{1n,x} + g_{2n,\tilde{y}} - n\pi\, g_{3n} = 0 \quad \text{for } n \ge 1.$$

As in Lemma 3.2, it is sufficient to consider (3.25)–(3.27) when $n \ge 1$. By taking Fourier transform in the $\tilde{y}$ direction, we obtain the algebraic linear system

$$\begin{aligned} -mk_1\,\widehat{u_{nm}} + i\,\xi\,\widehat{v_{nm}} + n\,\pi\,\widehat{w_{nm}} &= \widehat{f_{nm}}, \\ n\pi\,\widehat{v_{nm}} + i\,\xi\,\widehat{w_{nm}} &= \widehat{g^m_{1n}}, \\ -n\pi\,\widehat{u_{nm}} - mk_1\,\widehat{w_{nm}} &= \widehat{g^m_{2n}} \end{aligned} \tag{4.43}$$

for each $m \geq 1$. For $m, n \geq 1$, since $\rho < \rho_* < k_1$, the determinant $n^2\pi^2 + m^2 k_1{}^2 + \xi^2$of the above linear system has no complex roots in the strip $\mid \operatorname{Im} z \mid \leq \rho_*$. Now we can use the Cramer's rule and Lemma 4.3 to solve (4.43).

For $n = 0$ and $m \geq 1$, (3.25) and (3.28) lead to

$$-mk_1 \, \widehat{u_{0m}} + i\xi \, \widehat{v_{0m}} = \widehat{f_{0m}},$$
$$-i\xi \, \widehat{u_{0m}} + mk_1 \, \widehat{v_{0m}} = \widehat{g_{30}^{m}},$$

which can be solved as in the previous paragraph. When $n \geq 0$ and $m = 0$, from (3.28), we are led to solve

$$- i\xi \, \widehat{u_{n0}} = \widehat{g_{3n}^{0}}. \tag{4.44}$$

Note that (4.35) implies $\widehat{g_{3n}^{0}}(0) = 0$ for $n \geq 0$. Thus Lemma 4.3 applies and u_{n0} are determined.

With the help of the extra factor given in (4.36), we can show that

$$\mid u_n \mid_{s+1,\rho} + n \mid u_n \mid_{s,\rho} \leq C \left(|f_n|_{s,\rho} + \sum_{i=1}^{3} |g_{in}|_{s,\rho} \right),$$
$$\mid u_0 \mid_{s+1,\rho} + \mid v_0 \mid_{s+1,\rho} \leq C \mid (f, \vec{g}, 0) \mid_{Z_0}.$$

The rest of the proof is essentially the same as that of Lemma 5 in [1]. □

Now suppose that we are given a η of the form (4.27) in which η^0, $a^{\pm}$, and $\delta^{\pm}$ are within some bounded set. We can regard $J(\epsilon\eta)$ and $\{L_0 + \epsilon L_1(\eta)\}$ as operators, depending smoothly on η^0, $a^{\pm}$, $\delta^{\pm}$, and ϵ. Since $J(0)$ is the identity, $J(\epsilon\eta)$ is invertible for ϵ sufficiently small. Thus solving (4.30) is equivalent to solving

$$J(\epsilon\eta) \{ L_0 + \epsilon L_1(\eta) \} (u^0, v^0, w^0) = (f, \vec{g}, q), \tag{4.45}$$

where

$$(f, \vec{g}, q) = J(\epsilon\eta) \Big((0, \vec{0}, \eta_x) - \{ L_0 + \epsilon L_1(\eta) \} (u^{\pm}, v^{\pm}, w^{\pm}) \Big). \tag{4.46}$$

Here the last term with ± represents a sum of two terms. As a result of Lemma 4.5, $J(\epsilon\eta) \{L_0 + \epsilon L_1(\eta)\}$ is an invertible operator from Y onto Z_0, provided ϵ is sufficiently small. To solve (4.45), it remains to show that $(f, \vec{g}, q)$ defined above is in Z_0. Let

$$\{L_0 + \epsilon L_1(\eta)\}(u^{\pm}, v^{\pm}, w^{\pm}) = (f^{\pm}, g^{\pm}, q^{\pm}).$$

As before, if we apply (3.15) to $g^{\pm}$ and define $\tilde{g}^{\pm}$, then $\tilde{g}^{\pm}$ satisfies (4.32) and, as in (4.33), $\tilde{g}_3^{\pm}$ are divergences with respect to x and $\tilde{y}$. Thus $\vec{g} = -\tilde{g}^{+} - \tilde{g}^{-}$ defined in (4.46) satisfies (4.34). Moreover, from (4.30), we can see that $g^{\pm}$ defined above decay rapidly in the $\tilde{y}$ direction. This implies

$$\int_{-\ell_1}^{+\ell_1} g_3(x, \tilde{y}, z) \, dx = \partial_{\tilde{y}} \, r_1(\tilde{y}, z)$$

for some function r_1 that decays rapidly in the $\tilde{y}$ direction. Consequently, (4.35) and (4.36) hold and we have $(f, \vec{g}, q) \in Z_0$. Now, for ϵ sufficiently close to zero, (u^0, v^0, w^0) can be regarded as a function that depends smoothly on η^0, $a^{\pm}$, $\delta^{\pm}$, and ϵ through

$$(u^0, v^0, w^0) = \big(J(\epsilon\eta) \{ L_0 + \epsilon L_1(\eta) \} \big)^{-1} (f, \vec{g}, q), \tag{4.47}$$

where $(f, \vec{g}, q)$ is given in (4.46).

Now, because of (4.47), it remains to solve Bernoulli's equation (4.26) as a nonlinear equation of the form

$$B(\eta^0, a^\pm, \delta^\pm, \epsilon) + p = 0, \tag{4.48}$$

where

$$B(\eta^0, a^\pm, \delta^\pm, \epsilon) = \eta - \beta \tilde{\nabla} \cdot \frac{\tilde{\nabla}\eta}{[1+\epsilon^2 \mid \tilde{\nabla}\eta \mid^2]^{1/2}} + \gamma u + \frac{\epsilon}{2}(u^2+v^2+w^2). \tag{4.49}$$

Here $\tilde{\nabla} = (\partial_x,\, \theta\,\partial_{\tilde{y}})$; η and (u, v, w) are defined as in (4.27) and (4.28), with (u^0, v^0, w^0) given by (4.47). Let N be any bounded neighborhood of $\vec{0}$ in ${}_\rho H_e^{s+3/2}(\Gamma) \times \mathbf{R}^5$ and regard B as an operator defined on N. First, we point out that B ranges in ${}_\rho H_e^{s-1/2}(\Gamma)$. Let $(\eta^0, a^\pm, \delta^\pm, \epsilon)$ be an arbitrary element in N. Note that

$$\begin{aligned}\tilde{\nabla} \cdot \frac{\tilde{\nabla}\eta}{[1+\epsilon^2 \mid \tilde{\nabla}\eta \mid^2]^{1/2}} &= \frac{1}{\epsilon}\left\{\frac{\epsilon\,\eta_x}{[1+\epsilon^2\,\eta_x{}^2+\epsilon^2\,\theta^2\,\eta_{\tilde{y}}{}^2]^{1/2}}\right\}_x \\ &+ \frac{1}{\epsilon}\left\{\frac{\epsilon\,\theta\,\eta_{\tilde{y}}}{[1+\epsilon^2\,\eta_x{}^2+\epsilon^2\,\theta^2\,\eta_{\tilde{y}}{}^2]^{1/2}}\right\}_{\tilde{y}}.\end{aligned} \tag{4.50}$$

Let f be the C^∞ function

$$f(\tau_1, \tau_2) = \frac{\tau_1}{(1+\tau_1{}^2+\tau_2{}^2)^{1/2}}.$$

Note that

$$f(\tau_1+h_1, \tau_2+h_2) = f(\tau_1, \tau_2) + \int_0^1 \frac{d}{dt} f(\tau_1+th_1, \tau_2+th_2)\,dt.$$

Apply this formula with $\tau_1 = \epsilon\,(\eta_x^+ + \eta_x^-)$, $\tau_2 = \epsilon\,\theta\,(\eta_{\tilde{y}}^+ + \eta_{\tilde{y}}^-)$, $h_1 = \epsilon\,\eta_x^0$ and $h_2 = \epsilon\,\theta\,\eta_{\tilde{y}}^0$. Since $\eta^0 \in {}_\rho H_e^{s+3/2}(\Gamma)$, we have

$$f(\epsilon\,\eta_x, \epsilon\,\theta\,\eta_{\tilde{y}}) = f(\tau_1, \tau_2) + r_2,$$

where $r_2 \in {}_\rho H^{s+1/2}(\Gamma)$. By construction, the term

$$f(\tau_1, \tau_2) - f(\epsilon\,\eta_x^+, \epsilon\,\theta\,\eta_{\tilde{y}}^+) - f(\epsilon\,\eta_x^-, \epsilon\,\theta\,\eta_{\tilde{y}}^-)$$

has compact support in $\tilde{y}$. Hence, the first term in (4.50) can be rewritten as

$$\partial_x\, f(\epsilon\,\eta_x, \epsilon\,\theta\,\eta_{\tilde{y}}) = \partial_x\, f(\epsilon\,\eta_x^+, \epsilon\,\theta\,\eta_{\tilde{y}}^+) + \partial_x\, f(\epsilon\,\eta_x^-, \epsilon\,\theta\,\eta_{\tilde{y}}^-) + r_3,$$

where $r_3 \in {}_\rho H^{s-1/2}(\Gamma)$. We can rewrite the second term in (4.50) in a similar manner. Next, on $z = 1$, we can write

$$(u^2+v^2+w^2) = (u^{+2}+v^{+2}+w^{+2}) + (u^{-2}+v^{-2}+w^{-2}) + r_4,$$

where r_4 is in ${}_\rho H^{s+1/2}(\Gamma)$. Now, since the symmetric doubly periodic waves satisfy Bernoulli's equation, the sum of all the terms with ± sign has compact support in $\tilde{y}$. Hence, we have shown that B ranges in ${}_\rho H_e^{s-1/2}(\Gamma)$.

When $\epsilon = 0$, the y coordinate and the $\tilde{y}$ coordinate are the same since the stretching factor $\theta = 1$. Given phase shifts $\delta^\pm$ such that $\sin k_2(\delta^+ - \delta^-) \neq 0$, η^* and $a_0^\pm$ determined by (4.18) and (4.13) in §4.1 satisfy

$$B(\eta^*,\, a_0^\pm,\, \delta^\pm, 0) + p = 0.$$

Let N be a sufficiently large neighborhood of $(\eta^*, a_0^\pm, \delta^\pm, 0)$ in ${}_\rho H_e^{s+3/2}(\Gamma) \times \mathbf{R}^5$. In what follows, we will apply the implicit function theorem for Banach spaces to solve (4.48) for ϵ close to zero. We will show that the Fréchet derivative A of B with respect to $(\eta^0, a^\pm)$ at $(\eta^*, a_0^\pm, \delta^\pm, 0)$ is an invertible operator from ${}_\rho H_e^{s+3/2}(\Gamma) \times \mathbf{R}^2$ to ${}_\rho H_e^{s-1/2}(\Gamma)$.

When $\epsilon = 0$,

$$B(\eta^0, a^\pm, \delta^\pm, 0) = \eta - \beta \widetilde{\nabla}^2 \eta + \gamma\, u, \tag{4.51}$$

where η is of the form (4.27) and (u, v, w) is determined by (4.2)–(4.6). As is pointed out in (3.43), the terms $\eta^\pm$ in (4.27) now have the form

$$\eta^\pm = a^\pm\, \zeta^\pm(\tilde{y})\, \cos k_1 x\, \cos k_2(\tilde{y} - \delta^\pm)$$

when $\epsilon = 0$. Given any amplitudes $\tilde{a}^\pm$ and $\tilde{\eta}^0$ in${}_\rho H_e^{s+3/2}(\Gamma)$, with $\tilde{\eta}^0$ written as

$$\tilde{\eta}^0 = \sum_{m=0}^{+\infty} \tilde{\eta}_m(\tilde{y})\, \cos m k_1 x,$$

we obtain

$$\begin{aligned} \widehat{A(\tilde{\eta}^0,\, \tilde{a}^\pm)}(x, \xi) = \sum_{m=0}^{+\infty} b(m k_1, \xi)\, \widehat{(\tilde{\eta}_m)}(\xi)\, \cos m k_1 x \\ + \tilde{a}^+\, \partial_{a^+}\, b(k_1, \xi)\, \widehat{(\eta^+)}(x, \xi) \\ + \tilde{a}^-\, \partial_{a^-}\, b(k_1, \xi)\, \widehat{(\eta^-)}(x, \xi). \end{aligned} \tag{4.52}$$

Here $\widehat{\ }$ is the Fourier transform in $\tilde{y}$. We need to show that, for any $g \in {}_\rho H_e^{s-1/2}(\Gamma)$

$$A(\tilde{\eta}^0,\, \tilde{a}^\pm) = g \tag{4.53}$$

has a unique solution depending boundedly on g. Assume that

$$g = \sum_{m=0}^{+\infty} g_m(\tilde{y})\, \cos m k_1 x.$$

From (4.52), the transform of (4.53) leads to

$$\begin{aligned} \Big[b(k_1, \cdot)\, \widehat{(\tilde{\eta}_1)}\Big](\xi)\, \cos k_1 x = \widehat{g_1}(\xi)\, \cos k_1 x \\ - \tilde{a}^+\, \partial_{a^+}\, \Big[b(k_1, \cdot)\, \widehat{(\eta^+)}\Big](x, \xi) \\ - \tilde{a}^-\, \partial_{a^-}\, \Big[b(k_1, \cdot)\, \widehat{(\eta^-)}\Big](x, \xi), \end{aligned} \tag{4.54}$$

and

$$\left[b(mk_1, \cdot)\,\widehat{(\tilde{\eta}_m)} \right](\xi) = \widehat{g_m}(\xi) \quad \text{for } m \neq 1. \tag{4.55}$$

Note that (4.12) and (4.54) are similar. As before, we set $\xi = k_2$ and use

$$\tilde{a}^+ \left[b(k_1, \cdot)\,\widehat{(\eta^+)} \right](x, k_2) + \tilde{a}^- \left[b(k_1, \cdot)\,\widehat{(\eta^-)} \right](x, k_2) = \widehat{g_1}(k_2)\, \cos k_1 x$$

to form a linear system for $\tilde{a}^\pm$. Then by (4.11) and the assumption that $\sin k_2(\delta^+ - \delta^-)$ is nonzero, $\tilde{a}^\pm$ can be determined as in (4.13), and then the right-hand side of (4.54) becomes a function that vanishes at $\xi = k_2$. As in §4.1, the amplitudes $\tilde{a}^\pm$ thus determined guarantee that the right-hand side of (4.54) also vanishes at $\xi = -k_2$. As a result of Corollary 4.2, we can now use Lemma 4.3 to solve (4.54) and (4.55) as in (4.7), (4.15). The solution $(\tilde{\eta}^0, \tilde{a}^\pm)$ thus obtained can be estimated as in (4.14) and (4.19). We obtain, as in (4.21),

$$|\, \tilde{\eta}^0 \,|_{s+3/2,\rho} + |\, \tilde{a}^+ \,| + |\, \tilde{a}^- \,| \le C \,|\, g \,|_{s-1/2,\rho}.$$

The uniqueness of the solution follows as a result of the above estimate.

Now, since the Fréchet derivative A is invertible, the implicit function theorem for Banach spaces applies. As a result, there exists a solution $(\eta^0(\epsilon), a^\pm(\epsilon))$ of (4.48) in ${}_\rho H_e^{s+3/2}(\Gamma) \times \mathbf{R}^2$ for each ϵ sufficiently close to zero, provided the phase shifts at infinity $\delta^\pm$ satisfy $\sin k_2(\delta^+ - \delta^-) \neq 0$.

By examining the above result more closely, we can construct three-dimensional exact steady water waves traveling along a vertical wall located on $y = 0$, generated by a partially localized pressure disturbance. Suppose that the pressure disturbance $p \in {}_\rho H_e^{s-1/2}(\Gamma)$ is an even function of y. Observe that an L^2 function $v(x, y, z)$ is even in y if and only if its Fourier transform $\widehat{v}(x, \xi, z)$ is even in ξ. Similarly, v is odd in y if and only if $\widehat{v}$ is odd in ξ. Note that p is even in y implies that each of the $\widehat{p_m}(\xi)$ in (4.7) is even in ξ. Since the function $b(mk_1, \xi)$ is also even in ξ for each m, the solution η_m of (4.7) is even in y. For this reason, we seek a surface elevation η which is even in y. Given such a η, when we solve (4.2)–(4.6), the resulting (u, v, w) would have u, w even in y and v odd in y. Consequently, we will need to use Sobolev spaces which specify the symmetries in both the x and y direction. Let $H^s(\Omega)$ and $H^s(\Gamma)$ be defined as in §2. A subscript (e, e), (e, o), (o, e), or (o, o) is used to indicate the symmetries in x and y. Similarly, we define subspaces of ${}_\rho H^s(\Omega)$and ${}_\rho H^s(\Gamma)$. In what follows, we will only point out the necessary changes; most of the construction remains the same.

When we carry out the linear calculation as in §4.1, since η is even in y, we should take

$$a^+ = a^- \quad \text{and} \quad \delta^+ = -\delta^- \tag{4.56}$$

in (4.8). As a result, (4.13) is replaced by

$$2i\, a^+ \, \sin k_2 \delta^+ = \widehat{p_1}(k_2) \,/\, c, \tag{4.57}$$

where c is nonzero and pure imaginary. Since p_1 is real-valued and is even in y, we have

$$\overline{\widehat{p_1}(k_2)} = \widehat{p_1}(-k_2) = \widehat{p_1}(k_2).$$

This implies $\widehat{p_1}$ is real-valued. Given δ^+ such that $\sin k_2\delta^+ \neq 0$, we can determine a^+ uniquely from (4.57). The rest of the calculation remains unchanged except that we now have additional symmetry in the y direction. The estimate (4.21) shows that η^* is in ${}_\rho H^{s+3/2}_{e,e}(\Gamma)$. Because of (4.56), (4.17) provides a one-parameter family of solutions of (4.2)–(4.6) with v vanishing along $y = 0$.

As for modifying the construction in §4.2, first we need to replace function spaces Y and Z by

$$\begin{aligned} Y &= \{(u,v,w) \in {}_\rho H^{s+1}_{e,e}(\Omega) \times {}_\rho H^{s+1}_{o,o}(\Omega) \times {}_\rho H^{s+1}_{o,e}(\Omega) : w(x,\tilde{y},0) = 0\}, \\ Z &= {}_\rho H^{s}_{o,e}(\Omega) \times {}_\rho H^{2}_{o,o}(\Omega) \times {}_\rho H^{s}_{e,e}(\Omega) \times {}_\rho H^{s}_{e,o}(\Omega) \times {}_\rho H^{s+1/2}_{e,e}(\Gamma). \end{aligned}$$

Again, let Z_0 be the subspace of Z consisting of $(f, \vec{g}, q)$ that satisfies (4.34)–(4.36). Next, because of the symmetry in the $\tilde{y}$ direction, the function B in (4.49) can be regarded as a function of η^0, a^+, δ^+, and ϵ from a neighborhood N in ${}_\rho H^{s+3/2}_{e,e}(\Gamma) \times \mathbf{R}^3$ to ${}_\rho H^{s-1/2}_{e,e}(\Gamma)$. Amplitude a^- and phase shift δ^- on the other side are determined by (4.56). Consequently, when we solve (4.53), $\tilde{a}^+$ is determined by an equation that is similar to (4.57). Finally, by applying the implicit function theorem, we obtain a one-parameter family of exact solutions of (4.22)–(4.26). Note that these solutions also satisfy

$$v(x,0,z) = 0,$$

which is the boundary condition along the vertical wall. Now, if we accept the assumption that the contact angle of the fluid with the wall is 90° by restricting the solutions we just obtained to the domain $\{(x,y,z) : y > 0,\ 0 < z < 1\}$, we obtain a one-parameter family of exact steady waves traveling along a vertical wall.

5. Appendix.

Proof of Lemma 3.1. Let $f(x) = (1 + \beta x^2)\, x \tanh x$ for $x \geq 0$. From (3.3), it is easy to see that ξ is a real root of $b(k_1, \xi) = 0$ if and only if $\tilde{k} = ({k_1}^2 + \xi^2)^{1/2}$ satisfies

$$f(x) = \gamma\, {k_1}^2. \tag{5.1}$$

Thus we are led to investigate whether (5.1) has a positive root $\geq k_1$.

Clearly, we have $f'(x) > 0$ for any $x > 0$ and $f(0) = 0$. Hence, given any real number $k_1 > 0$, there exists a unique $\tilde{k} > 0$ such that $f(\tilde{k}) = \gamma\, {k_1}^2$. Note that

$$\begin{aligned} f(\tilde{k}) - f(k_1) &= \gamma\, {k_1}^2 - (1 + \beta\, {k_1}^2)\, k_1 \tanh k_1 \\ &= -k_1\, b(k_1, 0) \tanh k_1. \end{aligned} \tag{5.2}$$

From the proof of Lemma 2 in [1], it is not hard to see that

$$\begin{aligned} &b(k_1, 0) < 0 \quad \text{if } k_G < k_1 < k_T, \\ &b(k_1, 0) > 0 \quad \text{if } 0 < k_1 < k_G \ \text{ or } \ k_T < k_1. \end{aligned} \tag{5.3}$$

Thus, when $0 < k_1 < k_G$ or $k_T < k_1$ we have $f(\tilde{k}) < f(k_1)$. This implies that the only positive root of (5.1) is strictly less than k_1 and so $b(k_1, \xi) = 0$ has no real roots. When $k_G < k_1 < k_T$,(5.2) and (5.3) imply (5.1) has a positive root $\tilde{k}$ which is $> k_1$. As a result, $k_2 = (\tilde{k}^2 - {k_1}^2)^{1/2}$ satisfies $b(k_1, k_2) = 0$. The uniqueness of the positive root $\xi = k_2$ follows from that of $\tilde{k}$.

It remains to show that $k_2 > 0$ is a simple root of $b(k_1, \xi) = 0$. For any $\xi > 0$, note that

$$f\left(\sqrt{k_1{}^2 + \xi^2}\right) = \gamma\, k_1{}^2 + b(k_1, \xi)\sqrt{k_1{}^2 + \xi^2}\,\tanh\sqrt{k_1{}^2 + \xi^2}.$$

This implies

$$\left[\partial_\xi\, f\left(\sqrt{k_1{}^2 + \xi^2}\right)\right]_{\xi = k_2} = \left[\tilde{k}\,\tanh \tilde{k} \partial_\xi\, b(k_1, \xi)\right]_{\xi = k_2},$$

where $\tilde{k} = (k_1{}^2 + \xi^2)^{1/2}$. It is easy to see that $[\partial_\xi\, f(\sqrt{k_1{}^2 + \xi^2})]_{\xi = k_2}$ is nonzero. Thus $\partial_\xi\, b(k_1, \xi)$ is also nonzero at $\xi = k_2$. This completes the proof of Lemma 3.1. □

Proof of Corollary 4.2. Let ρ be an arbitrary constant such that $0 < \rho < k_1$. By Lemma 4.1, for each $m \geq 0$, $b(mk_1, z)$ is analytic over the strip $|\operatorname{Im} z| \leq \rho$. Moreover, there exists a positive integer M such that

$$|b(mk_1, z)| \geq C\,(m^2 k_1{}^2 + |\operatorname{Re} z|^2)^o$$

whenever $|\operatorname{Re} z|$ or m is larger than M. Hence, to prove Corollary 4.2, it suffices to consider analytic functions $b(0, z)$, $b(k_1, z), \ldots, b(Mk_1, z)$ over the compact set $\{\xi + i\phi : |\ \xi\ | \leq M,\ |\ \phi\ | \leq \rho\}$. Each of these functions can only have a finite number of zeros in the compact set. Since $\max(k_G, k_T/2) < k_1 < k_T$, among these functions, Lemma 3.1 implies that only $b(k_1, z)$ has two real roots $z = \pm k_2$. Thus there is a small enough $\rho_* > 0$ such that all the roots of $b(0, z)$, $b(k_1, z), \ldots, b(Mk_1, z)$ except $\pm k_2$ lie outside the strip $|\operatorname{Im} z| \leq \rho_* < k_1$. This proves Corollary 4.2. □

Acknowledgment. This paper is based on the results obtained in the author's dissertation. The author would like to thank his thesis advisor, Professor J. Thomas Beale, for his encouragement and advice. The author would also like to thank the referees for their helpful suggestions.

REFERENCES

[1] J. T. Beale, *Water waves generated by a pressure disturbance on a steady stream,* Duke Math. J., 47 (1980), pp. 297–323.

[2] D. E. Hewgill, J. Reeder, and M. Shinbrot, *Some exact solutions of the nonlinear problem of water waves,* Pacific J. Math., 92 (1981), pp. 87–109.

[3] J. Lighthill, *Waves in Fluids,* Cambridge University Press, Cambridge, 1978.

[4] P. I. Plotnikov, *Solvability of the problem of spatial gravitational waves on the surface of an ideal fluid,* Dokl. Akad. Nauk SSSR, 251 (1980), pp. 591–594.

[5] J. Reeder and M. Shinbrot, *Three-dimensional, nonlinear wave interactions in water of constant depth,* J. Nonlinear Anal. Theory Meth. Appl., 5 (1981), pp. 303–323.

[6] W. Rudin, *Real and Complex Analysis,* McGraw-Hill, New York, 1974.

[7] J. V. Wehausen and E. V. Laitone, *Surface waves,* in Handbuch der Physik IX, Springer-Verlag, Berlin, 1960, pp. 446–778.

[8] G. B. Whitham, *Linear and Nonlinear Waves,* John Wiley, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 24, No. 5, pp. 1179–1214, September 1993

004

TRAVELLING WAVES IN PREDATOR-PREY SYSTEMS*

KONSTANTIN MISCHAIKOW† AND JAMES F. REINECK‡

Abstract. The existence of travelling wave solutions to reaction-diffusion equations which model predator-prey systems is proven. Bistable waves, Fisher waves, and higher-dimensional analogues of Fisher waves are found. Some of the systems investigated have bistable homoclinic waves. The proofs use the Conley index, continuation, the connection matrix, and bifurcation theory in the Conley index setting.

Key words. travelling wave, reaction-diffusion equation, predator-prey system, Conley index

AMS subject classifications. primary 35K55; secondary 58E99

1. Introduction.

This paper discusses the existence of travelling wave solutions for systems of reaction-diffusion equations which model two-species predator-prey interactions. In particular, the following two questions are addressed.

(Q1) Given general qualitative hypotheses on the nonlinear reaction terms, what travelling waves occur?

(Q2) Given qualitative properties of a travelling wave, can one find nonlinearities for which such a travelling wave exists?

From the very beginning it must be emphasized that we are *far* from being able to give a complete answer to either of these questions. On the other hand, we are able to provide partial solutions to both problems. What is perhaps more important is that our results are obtained in a systematic manner which shows potential for applications in both higher-dimensional and more complicated problems.

In some sense the techniques we use are not new. They are based on careful choices of isolating neighborhoods and applications of the Conley index, a procedure which in a very simple setting (one-species) was outlined by Conley [Co, Chap. IV.2.6]. That these ideas work for more challenging problems has been amply demonstrated by now. For a sampling of these we draw the readers' attention to the work of Conley and Gardner [C-G] (a "simple" two-competing species model), Feinberg and Terman [F-T] (a "complicated" two-competing species model), Mischaikow and Hutson [M-H] (n-mutualist species), Mischaikow [Mi2] (two-mutualist and one-competitor), Gardner and Smoller [G-S] (periodic wave trains via perturbation methods). Of particular note for the problem we consider here is the work of Gardner [Ga], also on a predator-prey system.

Our proof of existence uses techniques similar to those of Conley [Co], Conley and Gardner [C-G], and Gardner [Ga]. As in the earlier work, we deform the original system to a simpler one where the existence of travelling waves is shown by making direct computations using the Conley index and related machinery (Theorem 2.1 and §4). A set is constructed in phase space which is an isolating neighborhood throughout the deformation, and it follows that there is a travelling wave throughout the deformation, and in particular for the original system. The major difference between

*Received by the editors May 16, 1991; accepted for publication (in revised form) January 5, 1993. This research was supported in part by the National Science Foundation.

†School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

‡Department of Mathematics, SUNY at Buffalo, Buffalo, New York, 14214.

our work and that of Conley–Gardner and Gardner is in the computations in the simplified system. We use connected simple system arguments instead of the connection index arguments of [C-G] and [Ga]. Thus we return to the ideas developed in Conley's monograph [Co]. Theorem 2.1 is weaker than the abstract theorem in [C-G] and [Ga], but it suffices to prove existence of travelling waves both in the earlier work and in our setting, and the necessary computations are simpler. In addition, we believe that this is an improvement over previous work, since, as will be demonstrated in §4, we can incorporate the connection matrix and transition matrix theory to obtain travelling waves in a product system which we then continue back to the original system. We believe that the continuation ideas developed in [Co], [C-G], and [Ga], combined with the more powerful computational tools which have been developed, will yield many useful results similar to those obtained in this paper.

Turning to the problem at hand, we will consider the following pair of reaction-diffusion equations

$$\begin{aligned} \frac{\partial u_1}{\partial t} &= \mu_1 \frac{\partial^2 u_1}{\partial x^2} + u_1 f(u_1, u_2), \\ \frac{\partial u_2}{\partial t} &= \mu_2 \frac{\partial^2 u_2}{\partial x^2} + u_2 g(u_1, u_2), \end{aligned} \tag{1.1}$$

where $x, t \in \mathbf{R}$ and the μ_i (assumed > 0) are the diffusion coefficients. It is convenient to state the hypothesis for (1.1) in terms of the *reaction system*

$$\begin{aligned} \dot{u}_1 &= u_1 f(u_1, u_2), \\ \dot{u}_2 &= u_2 g(u_1, u_2). \end{aligned} \tag{1.2}$$

H1. $f, g \in C^2$, $\partial f/\partial u_2 < 0$ and $\partial g/\partial u_1 > 0$.

H2. The zero sets of f and g, i.e., $\{u \mid f(u) = 0\}$ and $\{u \mid g(u) = 0\}$, are as shown in Fig. 1.1. In particular, $\{u \mid f(u) = 0\}$ is given by a smooth curve $p(u_1)$ lying to the right of the y-axis such that $p''(u_1) < 0$ and p has a unique maximum at $u_1 = \xi_1$. Similarly, $\{u \mid g(u) = 0\}$ is given by a smooth curve $q(u_2)$ lying to the right of the y-axis such that $q''(u_2) > 0$ and q has a unique minimum at $u_2 = \xi_2$. Furthermore, we assume that p and q intersect in four points with two intersections on each side of the vertex of either curve.

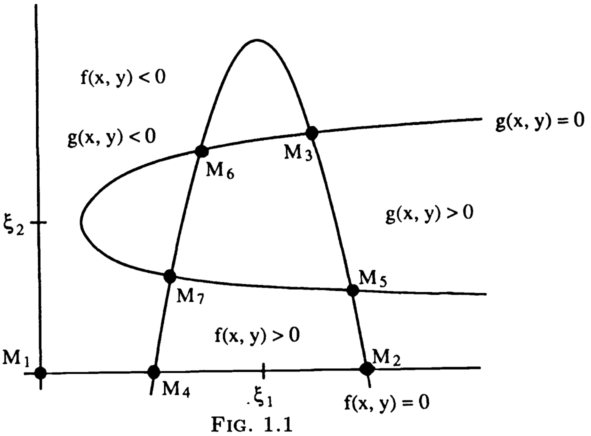


FIG. 1.1

From H1 and H2 it follows that there are seven critical points for (1.2). We label these as $M_i, i = 1, \ldots, 7$ and let (a_i, b_i) denote the coordinates of M_i.

H3. The critical points are hyperbolic.

This implies that M_1, M_2, and M_3 are attractors, M_4, M_5, and M_6 are saddle points, and M_7 is a repeller. Recall that a travelling wave for the system (1.1) is a solution of the form

$$u(x+\theta t) = (u_1(x+\theta t), u_2(x+\theta t)).$$

where θ is called the wave speed. If we set $\tau = x + \theta t$ and let $\dot{} = d/d\tau$, then (1.1) reduces to the four-dimensional system of ordinary differential equations,

$$\begin{aligned} \dot{u}_1 &= v_1, \\ \dot{u}_2 &= v_2, \\ \mu_1 \dot{v}_1 &= \theta v_1 - u_1 f(u_1, u_2), \\ \mu_2 \dot{v}_2 &= \theta v_2 - u_2 g(u_1, u_2). \end{aligned} \tag{1.3}$$

Notice that the critical points of (1.2) are in 1–1 correspondence with those of (1.3). More precisely, the set of critical points of (1.3) is given by $\{M(i) = (M_i, 0) \in \mathbf{R}^4 \mid i = 1, \dots, 7\}$.

Typically, when one speaks of travelling wave solutions one incorporates the boundary conditions

$$\begin{aligned} \lim_{\tau \to -\infty} (u(\tau), v(\tau)) &= (M_i, 0), \\ \lim_{\tau \to +\infty} (u(\tau), v(\tau)) &= (M_j, 0). \end{aligned} \tag{1.4}$$

We shall refer to such a solution as an $M(i) \to M(j)$ wave. In analogy with the one species problem (see Fife [Fi, pp. 106–109] for example) if $i, j = 1, 2, 3$ we shall refer to these waves as *bistable travelling waves.* From the point of view of stability (under the dynamics of the partial differential equation (1.1)) these solutions are of special interest.[1] Another class of waves which have been extensively studied are *Fisher waves.* For the specific system we are considering here these waves satisfy the boundary conditions (1.4) with $i = 4, 5, 6$ and $j = 1, 2, 3$. Because we are studying a system where the reaction dynamics is nontrivial, "new"[2] types of waves appear. From the point of view of the dynamics of the partial differential equation (1.1) it appears that the most relevant are those that satisfy the boundary conditions (1.4) with $i = 7$ and $j = 1, 2, 3$. We shall refer to these a HDF waves (higher-dimensional Fisher waves).

As will become clear in §8, we can, for high wave speeds, give a fairly complete description of the Fisher waves and HDF waves which occur under only slightly stronger hypotheses than H1–H3. Unfortunately, we cannot treat the bistable waves in this generality.

To motivate the additional hypotheses that we are forced to make consider the existence of an $M(1) \to M(2)$ wave. Observe that $\{(u_2, v_2) = (0, 0)\}$ defines a two–dimensional invariant subspace for (1.3) on which the dynamics are determined by

$$\begin{aligned} \dot{u}_1 &= v_1, \\ \mu_1 \dot{v}_1 &= \theta v_1 - u_1 f(u_1, 0). \end{aligned} \tag{1.5}$$

[1] This in no way implies that we claim to understand the stability properties of the waves whose existence we are proving. For techniques which might be applicable to some of our results, see [G-J].

[2] In this context new means that we have not seen it addressed in the literature.

Now it is easy to check that transformations $v_1 \mapsto -v_1$, $\theta \mapsto -\theta$, and $\tau \mapsto -\tau$ leaves (1.5) unchanged (in fact it leaves (1.3) unchanged). Thus there is no loss of generality in restricting our attention to $\theta \geq 0$. To simplify the presentation we shall actually search for waves that satisfy the slightly stronger assumption condition that $\theta > 0$.

Define

$$H(u_1, v_1) = \frac{\mu_1}{2} v_1^2 + \int u_1 f(u_1, 0)\, du_1. \tag{1.6}$$

Then along solutions to (1.5),

$$\frac{dH}{d\tau} = \theta v_1^2 \geq 0. \tag{1.7}$$

Thus, a necessary condition for the existence of an $M(1) \to M(2)$ wave is the assumption

$$\int_{a_1}^{a_2} u_1 f(u_1, 0)\, du_1 > 0. \tag{H4}$$

This is also a sufficient condition as the following theorem indicates.

THEOREM 1.1. *Given* H1–H4, *there exists an* $M(1) \to M(2)$ *wave for some wave speed* $\theta_{12} > 0$.

This is by now a classical result (see Fife [Fi], Conley–Gardner [C-G], or Terman [Te], for instance). For a proof in the spirit of the rest of the results of this paper see Conley [Co]. Since it is obvious how to modify the assumptions of this theorem to prove the existence of an $M(2) \to M(1)$ which lies in the $\{u_2 = v_2 = 0\}$ invariant plane we shall not explicitly discuss this problem. This comment holds throughout the rest of the discussion.

Remark. Let $(u(t), v(t))$ be a travelling wave solution, and assume that a maximal interval of monotonicity for $u_1(t)$ is $[t_0, t_1]$. Then we immediately obtain

$$0 = \left.\frac{\mu_1 v_1^2(t)}{2}\right|_{t_0}^{t_1} = \mu_1 \int_{t_0}^{t_1} v_1 \dot{v}_1\, d\tau = \int_{t_0}^{t_1} \theta v_1^2\, d\tau - \int_{t_0}^{t_1} u_1 v_1 f(u_1, u_2)\, d\tau.$$

Using the monotonicity, we observe that this integral can be parameterized by u_1, i.e.,

$$0 = \mu_1 \int_{t_0}^{t_1} \theta v_1^2\, d\tau - \int_{\alpha}^{\beta} u_1 f(u_1, u_2(u_1))\, du_1,$$

where $\alpha = u_1(t_0)$ and $\beta = u_1(t_1)$. For example, using the wave $M(1) \to M(2)$, this reduces to

$$\theta = \frac{\int_{a_1}^{a_2} u_1 f(u_1, 0)\, du_1}{\mu_1 \int_{-\infty}^{\infty} \theta v_1^2\, d\tau},$$

and H4 clearly implies that

$$\theta > 0.$$

Although this argument is trivial, it will be repeated numerous times in various forms to guarantee that the waves considered have positive wave speed.

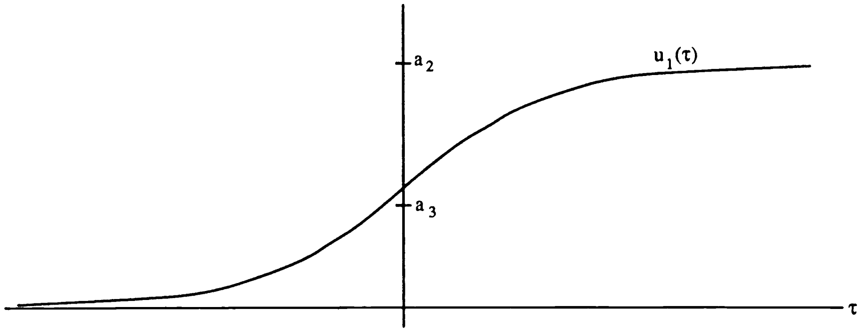


FIG. 1.2. $M(1) \to M(2)$ *wave from Theorem* 1.1.

It is enlightening, perhaps, to consider what the $M(1) \to M(2)$ waves of Theorem 1.5 look like in terms of the function $u_1(\tau)$. See Fig. 1.2.

We now turn to the first new result, namely the existence of an $M(2) \to M(3)$ wave. In this case there does not exist a simple two-dimensional invariant subspace, and hence we are forced to work with (1.3). The strategy in this case involves the following three steps.

1. Homotope the four-dimensional system to a system with an invariant two-dimensional subsystem on which the dynamics are qualitatively the same as (1.5).

2. Use the proof of Theorem 1 to conclude the existence of a wave speed for which the heteroclinic orbit occurs for the two-dimensional system.

3. Homotope back to the original problem and conclude that for some wave speed the travelling wave exists.

Since we will be using Conley index techniques, the key to this procedure is the existence of an isolating neighborhood throughout the homotopy. The following hypotheses will be used to guarantee this. Let α and γ be defined by Fig. 1.3. We assume

$$\text{(H5)} \qquad p(\xi_1) > \alpha,$$

$$\text{(H6)} \qquad \int_0^\gamma sg(a_2, s)\, ds < 0.$$

Finally, we add a hypothesis to guarantee that the wave speed is positive:

$$\text{(H7)} \qquad \int_0^{b_3} sg(a_3, s)\, ds > 0.$$

THEOREM 1.2. *Given* H1–H3, H5–H7, *there exists an* $M(2) \to M(3)$ *wave with wave speed* $\theta_{23} > 0$.

In Fig. 1.4 we indicate what the $M(2) \to M(3)$ wave of Theorem 1.2 looks like. Note that there are two possibilities: (a) where the eigenvalues of $(M_3, 0)$ are real and (b) where the eigenvalues are complex.

The next wave whose existence will be demonstrated is $M(3) \to M(1)$. This result is complicated by the fact that there does not appear to be a natural two-dimensional system towards which one can homotope. Thus instead we will homotope to a simple four-dimensional system for which the nonlinearities decouple. Again we

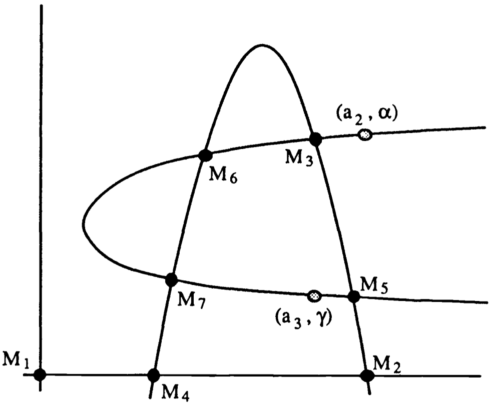


FIG. 1.3

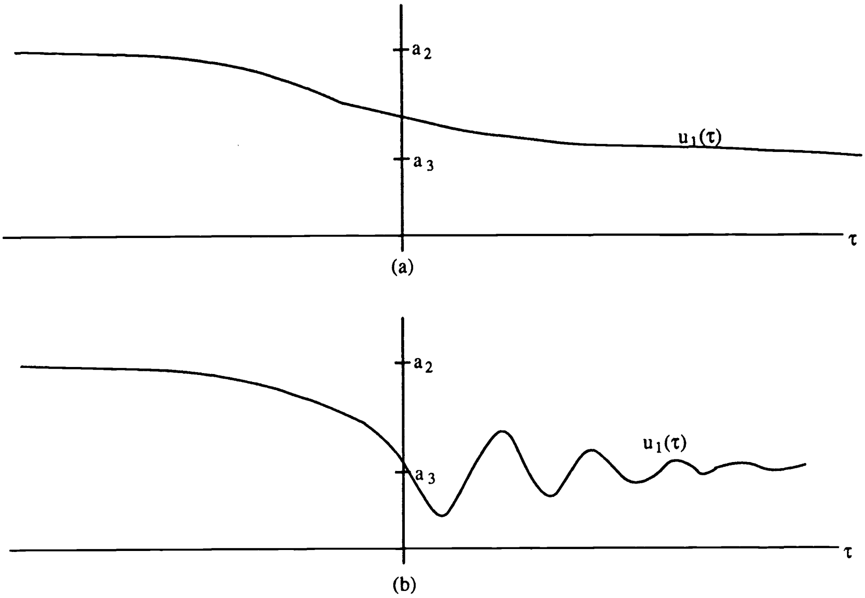


FIG. 1.4. $M(2) \to M(3)$ *wave from Theorem* 1.2.

need hypotheses to guarantee the existence of isolating neighborhoods and a positive wave speed.

$$\text{(H8)} \qquad \int_{a_4}^{a_5} sf(s, b_3)\, ds > 0,$$

$$\text{(H9)} \qquad \int_{0}^{b_3} sg(a_4, s)\, ds > 0.$$

THEOREM 1.3. *Given* H1–H3, H8, *and* H9, *there exists an* $M(3) \to M(1)$ *wave for some wave speed* $\theta_{31} > 0$.

Figure 1.5 indicates what the function $u_1(\tau)$ looks like for the $M(3) \to M(1)$ wave of Theorem 1.3. Again there are two possibilities depending on whether the eigenvalues of $(M_3, 0)$ are real (a) or complex (b).

Theorems 1.1, 1.2, and 1.3 are partial answers to question Q1. We shall now describe our results involving Q2.

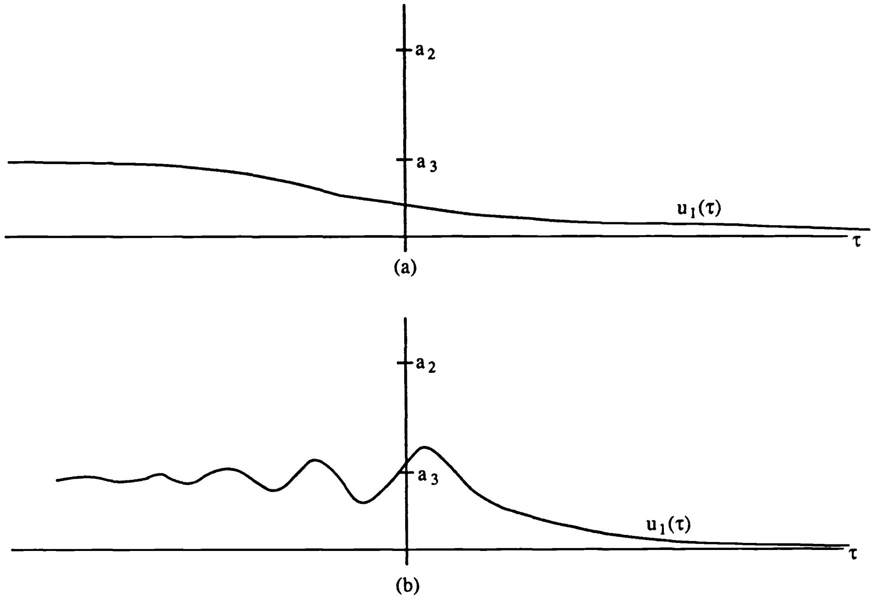


FIG. 1.5. *The* $M(3) \rightarrow M(1)$ *wave of Theorem* 1.3.

We begin by viewing (1.1) as a parameterized family of equations, namely

$$\begin{aligned}\frac{\partial u_1}{\partial t} &= \mu_1 \frac{\partial^2 u_1}{\partial x^2} + u_1 f^\lambda(u_1, u_2),\\ \frac{\partial u_2}{\partial t} &= \mu_2 \frac{\partial^2 u_2}{\partial x^2} + u_2 g^\lambda(u_1, u_2),\end{aligned} \tag{1.8}$$

where $\lambda \in [0,1]$. Furthermore, we assume that the hypothesis H1–H9 are satisfied for all values of λ. In this case Theorems 1.1, 1.2, and 1.3 apply and hence for each $\lambda \in [0,1]$ there exist positive wave speeds, θ^λ_{12}, θ^λ_{23}, and θ^λ_{31}, for which corresponding travelling waves occur. We would like to show that for some choice of nonlinearities additional bistable waves occur; e.g., $M(1) \rightarrow M(3)$, $M(2) \rightarrow M(1)$, $M(3) \rightarrow M(2)$, also with positive wave speed.

Unfortunately, at this point an additional complication arises, namely, the possibility that the wave speeds are not unique. It is well known that the $M(1) \rightarrow M(2)$ waves which satisfy (1.5) occur at a unique positive wave speed θ^λ_{12}, however, there are no corresponding results known regarding θ^λ_{23} and θ^λ_{31}. For each fixed λ, let $\bar{\theta}^\lambda_{23}$ denote the fastest wave speed and $\underline{\theta}^\lambda_{23}$ the slowest wave speed for which an $M(2) \rightarrow M(3)$ travelling wave occurs for (1.8^λ). By our assumptions $0 < \underline{\theta}^\lambda_{23} \leq \bar{\theta}^\lambda_{23} < \infty$. Let $I^\lambda_{23} = [\underline{\theta}^\lambda_{23}, \bar{\theta}^\lambda_{23}] \subset (0,\infty)$ and $I^\lambda_{31} = [\underline{\theta}^\lambda_{31}, \bar{\theta}^\lambda_{31}]$.

THEOREM 1.4. *Assume that for* $\lambda = 0$, *and* $i \neq j \neq k \neq i$, $\underline{\theta}^0_{ik} > \bar{\theta}^0_{kj}$ *and for* $\lambda = 1$, $\bar{\theta}^1_{ik} < \underline{\theta}^1_{kj}$, *there then exists an open interval* $\Lambda \subset [0,1]$, *such that for* $\lambda \in \Lambda$, *there exists an* $M(i) \rightarrow M(j)$ *wave which passes close to* $(M_k, 0)$ *and which has wave speed* $\theta^\lambda_{ij} > 0$.

It should be noted that the proof of this theorem works equally well if we assume that $\bar{\theta}^0_{ik} < \underline{\theta}^0_{kj}$ and $\underline{\theta}^1_{ik} > \bar{\theta}^1_{kj}$.

The implications of this theorem is most easily understood via Fig. 1.6, where the case $i = 2$, $j = 1$, and $k = 3$ is shown. Again there are two cases depending on the nature of the eigenvalues at $(M_3, 0)$. These drawings should be contrasted with that of Fig. 1.2.

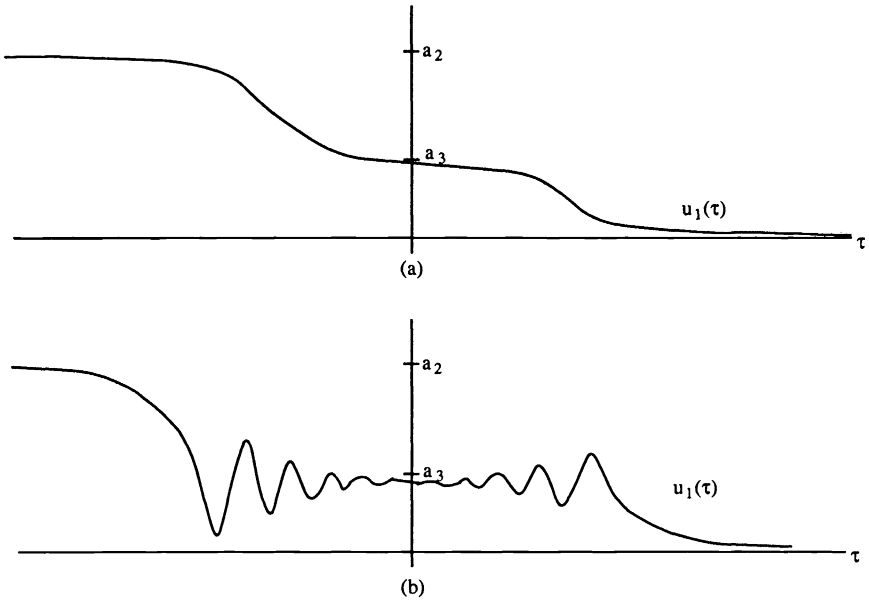


FIG. 1.6. *$M(2) \to M(1)$ wave of Theorem* 1.4.

Finally, we shall prove that for some set of nonlinearities there exist homoclinic pulse waves. Recall the notation of Theorem 1.4. In particular, for $\lambda \in \Lambda$, let θ_{ij}^{λ} denote the wave speed of the $M(i) \to M(j)$ wave of Theorem 1.4 and let θ_{ji}^{λ} denote the wave speed of the $M(j) \to M(i)$ wave of Theorems 1.1, 1.2, and 1.3.

THEOREM 1.5. *Let the interval $[a,b] \subset \Lambda$, and assume that $\underline{\theta}_{ij}^{a} > \bar{\theta}_{ji}^{a}$ and $\bar{\theta}_{ij}^{b} < \underline{\theta}_{ji}^{b}$. Then there exists an open interval $\widetilde{\Lambda} \subset \Lambda$, such that for $\lambda \in \widetilde{\Lambda}$ there exists an $M(i) \to M(i)$ wave which passes close to $(M_j, 0)$ and $(M_k, 0)$ and which has wave speed $\theta_{ii}^{\lambda} > 0$.*

Again, the implications of this result are explained in Fig. 1.7 (where $i = 2$), and depends on the eigenvalues at $(M_3, 0)$.

A large part of this paper is concerned with proving Theorems 1.2–1.5. In §2, the relevant definitions and theorems from the Conley index theory are presented. The proofs of Theorems 1.2 and 1.3 appear in §3 and §5, respectively. However, the latter result depends on understanding a complicated homotopy from a product system to the desired system. This is discussed in §4. In §6 the proofs of Theorems 1.4 and 1.5 are presented. This is followed in the next section by a construction to show that the abstract algebraic results of §6 are realizable for our particular class of predator-prey systems. We include a discussion in an attempt to convince the reader that these techniques are, in fact, applicable to a wide variety of travelling wave problems.

In §8 we return to the question of Fisher waves and HDF waves. The key to the existence proofs are two abstract results, Theorems 8.1 and 8.7. Since the theorems make it trivial to determine the existence of certain waves, the emphasis of this section is on explaining how they can be applied rather than formal existence theorems for the predator-prey system. Finally in §9 the proofs of Theorem 8.1 and Corollary 8.2 are presented.

2. Definitions and notation.

This section presents some of the definitions, notations, ideas, and theorems related to the Conley index theory. However, it is assumed that the reader is already familiar with these ideas (see Conley [Co], Salamon

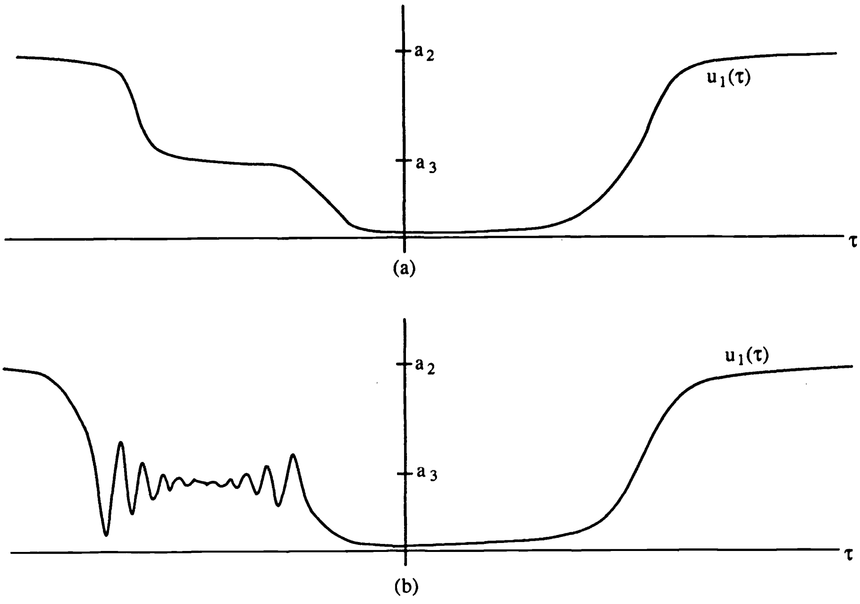


FIG. 1.7. *$M(2) \to M(2)$ wave of Theorem* 1.5.

[Sa], Smoller [Sm], Franzosa [Fr], or Moeckel [Mo]).

Let $\phi : \mathbf{R} \times X \to X$ be a flow on a locally compact space. Given a compact set N, let

$$N^T = \{x \in N \mid \phi([-T, T], x) \subset N\}.$$

N is called an *isolating neighborhood* if there exists $T > 0$ such that $N^T \subset \operatorname{int}(N)$. Let $N^\infty = \bigcap_{T>0} N^T$. It is easy to check that N^∞ is an invariant set, i.e., $\phi(\mathbf{R}, N^\infty) = N^\infty$. Clearly, N is an isolating neighborhood if and only if $x \in N^\infty$ implies $x \notin \partial N$. An invariant set S is called *isolated* if there exists an isolating neighborhood N of S such that $S = N^\infty$. We shall also use $I(N)$ to denote N^∞.

A simple but, as we shall see, useful decomposition of an invariant set S is that of an *attractor-repeller* (A-R) pair. Let $\omega(U)$ and $\omega^*(U)$ denote the omega and alpha limit sets of U, respectively. $A \subset S$ is called an *attractor* in S if there exists a neighborhood U of A such that $\omega(U \cap S) = A$. The *dual repeller* of A, denoted by A^* is defined by $A^* = \{x \in S \mid \omega(x) \cap A = \emptyset\}$. The pair (A, A^*) make up an A-R pair. Notice that given an A-R pair decomposition of S, if $x \in S$ then $x \in A \cup A^*$ or $\omega(x) \subset A$ and $\omega^*(x) \subset A^*$, i.e., S is made up of the attractor A, its dual repeller A^*, and connecting orbits from A^* to A. Thus, if one lets $C(A^*, A)$ denote the set of connecting orbits from A^* to A, then $S = A \cup A^* \cup C(A^*, A)$.

Since in our proof the complicated four-dimensional problem will be homotoped to a simple two- or four-dimensional problem we need to discuss what is meant by continuing isolating neighborhoods and isolated invariant sets. Consider a parameterized family of flows $\phi(t, x, \sigma) = \phi^\sigma(t, x)$, where $\sigma \in [0, 1]$ is the parameter value. Define the parameter flow to be $\Phi : \mathbf{R} \times X \times [0, 1] \to X \times [0, 1]$ by

$$\Phi(t, x, \sigma) = (\phi^\sigma(t, x), \sigma).$$

Let N^σ denote an isolating neighborhood for the flow ϕ^σ. One says that the isolating neighborhood N^1 continues to N^0 if there exists N an isolating neighborhood of Φ, the parameter flow, such that $N|_{X \times \{0\}} = N^0$ and $N|_{X \times \{1\}} = N^1$. Similarly, if S^σ denotes an isolated invariant set under ϕ^σ, then S^0 continues to S^1 if there exist corresponding isolating neighborhoods which continue.

For our applications, the homotopy we have in mind takes the form

$$(TW^{\theta,\sigma}) \qquad \begin{aligned} \dot{u}_1 &= v_1, \\ \dot{u}_2 &= v_2, \\ \mu_1^\sigma \dot{v}_1 &= \theta v_1 - u_1 f^\sigma(u_1, u_2), \\ \mu_2^\sigma \dot{v}_2 &= \theta v_2 - u_2 g^\sigma(u_1, u_2), \end{aligned}$$

where $\sigma \in [0,1]$ is the homotopy parameter. It will always be assumed that at $\sigma = 0$ we have the system of interest, namely, (1.3), and at $\sigma = 1$ we have a simpler system, i.e. one which we can analyze directly. The corresponding family of reaction systems is given by

$$(R^\sigma) \qquad \begin{aligned} \dot{u}_1 &= u_1 f^\sigma(u_1, u_2), \\ \dot{u}_2 &= u_2 g^\sigma(u_1, u_2). \end{aligned}$$

We now state an abstract theorem which will be used in §3 to prove Theorem 1.2, and a modification of which shall be used in §5 to prove Theorem 1.3.

Assumption 1. For the reaction system (R^1) there exists a one-dimensional attracting invariant affine subspace $L \subset \mathbf{R}^n$. Furthermore, the dynamics on L are as in Fig. 2.1, i.e., the set of bounded solutions consist of the hyperbolic critical points $\{A, B, C\}$ and the heteroclinic connections $C \to A$ and $C \to B$.

We now state the assumption concerning the homotopy from $\sigma = 1$ to $\sigma = 0$. Let $A_0^1 = (A^1, 0), B_0^1 = (B^1, 0)$, and $C_0^1 = (C^1, 0)$.

Assumption 2. A^σ and B^σ continue as critical points for (R^σ) and N^σ continue as isolating neighborhoods for $(TW^{c,\sigma})$ such that (B_0^σ, A_0^σ) is an attractor-repeller pair for $(N^{c,\sigma})^\infty$, the invariant set isolated by N^σ under the flow generated by $(TW^{c,\sigma})$. Furthermore, for $\sigma = 1$, if there exists an $A_0^1 \to B_0^1$ solution to $(TW^{c,1})$ for some value of c, then the connecting orbit belongs to $(N^{c,1})^\infty$.

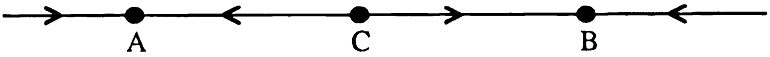


FIG. 2.1 *The dynamics on L.*

THEOREM 2.1 [M-H, Thm. 3.1]. *Given Assumptions* 1 *and* 2, *there exists a bistable travelling wave from* A_0 *to* B_0 *for some wave speed* θ_{AB}.

As will be seen most of §3 is dedicated to verifying Assumption 2.

Assumption 1 is equivalent to finding an invariant two-dimensional subspace for $(TW^{\theta,1})$ on which the dynamics is qualitatively equivalent to that of (1.5). For proving the existence of the $M(3) \to M(1)$ wave, this does not appear to be possible. Thus we need a weaker version of Assumption 1. This will be discussed in §4 and requires connection matrix and transition matrix techniques. For these purposes we shall adopt the following conventions.

The $\mathbf{Z}_2$ homology Conley index of S is denoted and defined by

$$CH_*(S) := H_*(N/L, [L]; \mathbf{Z}_2),$$

where (N, L) is an index pair for S. Given a Morse decomposition of S, $\mathbf{M}(S) = \{M(i) \mid i \in (P, >)\}$, where P is the indexing set for the Morse sets $M(i)$ and $>$ is a strict partial order on P, recall that the connection matrix is a linear map

$$\Delta : \bigoplus_{i \in P} CH_*(M(i)) \longrightarrow \bigoplus_{i \in P} CH_*(M(i)).$$

We shall always take the direct sum according to a nonincreasing order in P. This has the effect of making Δ into a strictly upper triangular matrix.

For (N, L) an index pair for S, the homotopy type of the pointed space $(N/L, [L])$ is usually referred to as the Conley index of S and is denoted $h(S)$. There exists, however, a finer version of the index which will be used in the proof. A *connected simple system* consists of a collection I_o of pointed spaces along with a collection I_m of homotopy classes of maps between these such that:

1. $\hom(X, X') = \{[f] \in [X, X'] \mid [f] \in I_m\}$ is nonempty and consists of a single element for each ordered pair X, X' of spaces in I_o;

2. if $X, X', X'' \in I_o, [f] \in \hom(X, X')$, and $[f'] \in \hom(X', X'')$, then $[f' \circ f] \in \hom(X, X'')$;

3. $\hom(X, X) = \{[1_X]\}$ for all $X \in I_m$.

Recall ([Co], [Sa]) that the Conley index of S forms a connected simple system where $I_o = \{(N/L, [L]) \mid (N, L) \text{ is an index pair for } S\}$ and I_m consists of the flow defined maps between the elements of I_o. The connected simple system of the Conley index of S is denoted by $I(S)$. The following result ([Co], [Sa]) is crucial to our analysis.

PROPOSITION 2.2. *If (A^c, A^{c*}) is an attractor repeller pair for S^c which continues for $c \in \mathbf{R}$ and $S^{c_i} = A^{c_i} \cup A^{c_i*}$ when $i = 0, 1$, but $I(S^{c_0}) \not\cong I(S^{c_1})$, then for some $c \in (c_0, c_1)$ there exists a connecting orbit from A^{c*} to A^c.*

It should be noted that the proof of Theorem 2.1 uses this proposition. In fact, a cursory reading of the proof [M-H, Appen.] shows that if the bistable wave is found via Theorem 2.1 then the connected simple systems at large and small wave speed differ.

3. The $M(2) \to M(3)$ travelling wave.

In this section we prove Theorem 1.2, i.e., we show that there is a $M(2) \to M(3)$ travelling wave provided assumptions H1–H3 and H5–H7 are satisfied. Recall that α and γ are defined in Fig. 1.3. Choose $l > \alpha$ such that $f(\xi_1, \alpha) > 0$ (we can do this by H5). Choose $h > a_2$ such that $\int_0^\gamma sg(h, s)\, ds < 0$ and such that $\{\, (x, l) \mid \xi_1 \le x \le h \,\}$ lies in the region $\{g < 0\}$. We can always do this by the assumptions on g. Choose k and m such that $\{\, (x, k) \mid \xi_1 \le x \le h \,\}$ lies in the region $\{g > 0\}$, and $\{\, (x, y) \mid \xi_1 \le x \le h, 0 \le y \le m \,\}$ lies in the region $\{g < 0\}$.

To prove Theorem 1.2, we use Theorem 2.1 and continue to a system where g is unchanged and $f(x, y)$ is independent of y. See Fig. 3.1. Clearly one can choose a deformation so that the assumptions hold for $(TW^{\theta,\sigma})$ for $\sigma \in [0, 1)$. In particular, the integral in assumption H7 can be bounded throughout the deformation, so by the Remark following Theorem 1.1 the wave speed will be bounded away from zero. For each σ we will construct an isolating neighborhood which contains only the critical points $M(2)$ and $M(3)$, plus any connecting orbits between the points. The construction we make will also construct an isolating neighborhood for $\sigma = 1$, so we can apply Theorem 2.1 and conclude the existence of an $M(2) \to M(3)$ travelling wave.

Let $N_3 = [\xi_1, h] \times [k, l]$, $N_4 = [\xi_1, h] \times [m, l]$, and $N_5 = [\xi_1, h] \times [-e, m]$ for small positive e. These will be the u projections of our isolating neighborhood. For $K, \epsilon > 0$,

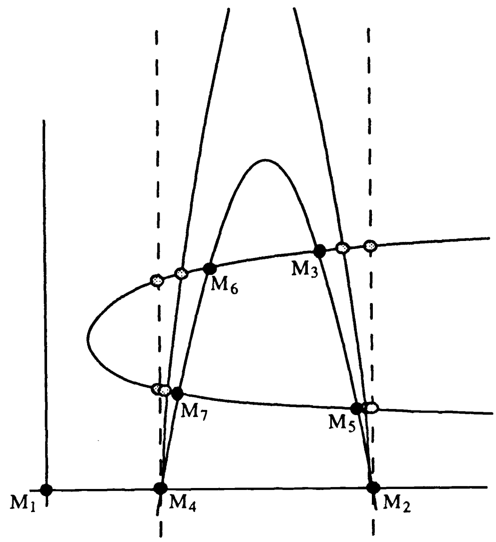


FIG. 3.1

define

$$\bar{N}_3 = N_3 \times [-K, K] \times [-K, K],$$

$$\bar{N}_4 = N_4 \times [-K, 0] \times [0, K],$$

$$\bar{N}_5 = N_5 \times [-K, K] \times [-K, K],$$

$$\bar{N}_{K,\epsilon} = (\bar{N}_3 \cup \bar{N}_4 \cup \bar{N}_5) \setminus B_\epsilon(M(5)).$$

See Fig. 3.2. Via a series of lemmas we will show that for K big enough and ϵ small enough, $\bar{N}_{K,\epsilon}$ is an isolating neighborhood.

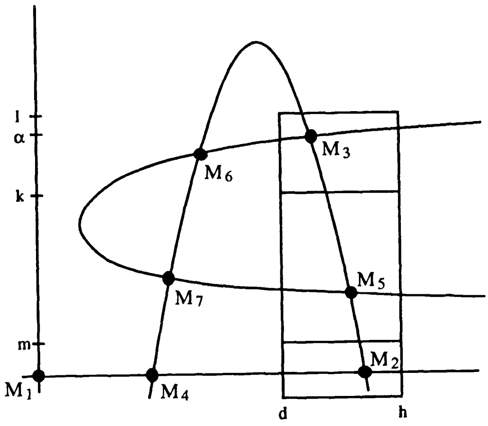


FIG. 3.2

LEMMA 3.1. *For any K, $I(\bar{N}_3) = M(3)$.*

Proof. Suppose there is some nonconstant orbit which stays in $\bar{N}_3$ for all time. Let $B = \{ (u_1, u_2) \mid x_1 \leq u_1 \leq x_2,\ y_1 \leq u_1 \leq y_2 \}$ be the smallest rectangular box containing the u-projection of the orbit. Since B is the smallest such box, somewhere on the top edge there is a point (u_1, y_2) such that when the u-projection of the orbit passes through the point, $v_2 = 0$ and $\dot{v}_2 \leq 0$. At this point we must have $-y_2 g(u_1, y_2) \leq 0$, i.e., $g(u_1, y_2) \geq 0$. It follows that $g(x_2, y_2) \geq 0$. By looking at the other edges of B, we can conclude that $f(x_2, y_1) \geq 0$, $g(x_1, y_1) \leq 0$, and $f(x_1, y_2) \geq 0$. Assumption H5 plus the concavity assumptions on p and q make it impossible for such a closed box to exist, so there is no recurrent orbit, and the only orbit which stays in $\bar{N}_3$ for all time is the critical point $M(5)$. □

LEMMA 3.2. *For any K, $I(\bar{N}_5) = M(2)$.*

Proof. We first show $I(N_5 \times [-K, K] \times [-K, K])$ is contained in the invariant plane $\{u_2 = v_2 = 0\}$. If $v_2 = 0$, then $\mu_2 \dot{v}_2 = -u_2 g(u)$, and since $g(u) < 0$ in N_5, the sign of $\dot{v}_2$ is the same as the sign of u_2 when $v_2 = 0$. Thus if $u_2 > 0$, and $v_2 > 0$, v_2 cannot change sign along the forward orbit and the orbit leaves $\bar{N}_5$ in forward time. If $u_2 > 0$ and $v_2 < 0$, then the orbit leaves $\bar{N}_5$ in backward time. If $u_2 < 0$, then a similar argument shows that the u-projection of the orbit leaves the bottom of N_5 in one direction or the other. Finally, if $u_2 = 0$ and $v_2 \neq 0$, then there are points on the orbit with $u_2 \neq 0$, so the orbit must leave. □

Notice that the proof of Lemma 3.2 actually shows that for an $M(2)$ to $M(3)$ connecting orbit, $v_2 > 0$ in $N_5 \times [-K, K] \times [-K, K]$. By Lemmas 3.1 and 3.2, and the fact that u_1 acts as a Lyapunov function in the region with u-projection in N_4, the only orbits which can stay in $\bar{N}_{K,\epsilon}$ for all time are connecting orbits from $M(2)$ to $M(3)$. To show that $\bar{N}_{K,\epsilon}$ is an isolating neighborhood, we must show that no *connecting* orbit is internally tangent to $\bar{N}_{K,\epsilon}$. For sufficiently large K, if $v_i = K$, then the u-projection of the orbit leaves $N_3 \cup N_4 \cup N_5$, so no internal tangencies can occur at $v_i = K$. Also, no orbit can be internally tangent to the top of N_3, the bottom of N_5, or the left or right side of the N_i. If $u_2 = l$, for example, and $v_2 \neq 0$, then the orbit leaves immediately. If $v_2 = 0$, then $\mu_2 \dot{v}_2 = -u_2 g(u) > 0$ so the orbit is externally tangent. The argument which shows that orbits tangent to the bottom and sides are externally tangent in u-space is similar.

LEMMA 3.3. *If (u, v) is a point on a connecting orbit in $\bar{N}_{K,\epsilon}$ and $u \in N_4$, then $v_i \neq 0$ for $i = 1, 2$.*

Proof. First assume that $u \in \operatorname{int}(N_4)$. If $v_1 = 0$ and $\dot{v}_1 \neq 0$, then the orbit leaves $\bar{N}_K$ immediately in one direction or the other. If $v_1 = 0$ and $\dot{v}_1 = 0$, then $-u_1 f(u) = 0$, and since $u_1 > 0$ in $\bar{N}_K$, we must have $f(u) = 0$. Thus if (u, v) is on a connecting orbit and $v_1 = \dot{v}_1 = 0$, then u must lie on the segment of $\{f = 0\}$ between M_2 and M_3. At this point we have

$$\mu_1 \ddot{v}_1 = \theta \dot{v}_1 - v_1 f(u) - u_1 v_1 f_1(u) - u_1 v_2 f_2(u) = -u_1 v_2 f_2(u).$$

(Here f_1 denotes the partial derivative with respect to the first argument.) If $v_2 > 0$, then $\ddot{v}_1 > 0$ and the orbit leaves in forward time. If $v_2 = 0$, then $\dot{v}_2 \neq 0$ (or else we're at a critical point), so the orbit leaves in one direction or the other. Thus v_1 cannot be zero.

If $v_2 = 0$ and $\dot{v}_2 \neq 0$, then the orbit leaves immediately in one direction or the other as noted above. If $\dot{v}_2 = 0$, then $g(u) = 0$. Parameterize time so that

$(u, v) = (u(0), v(0))$. Since $v_2 \geq 0$ we can parameterize the orbit by u_2. We get

$$\begin{aligned} 0 = \mu_2 v_2^2(0)/2 = \mu_2 \int_{-\infty}^{0} v_2 \dot{v}_2 \, d\tau &= \int_{-\infty}^{0} \theta v_2^2 \, d\tau - \int_{-\infty}^{0} u_2 v_2 g(u_1, u_2) \, d\tau \\ &= \int_{-\infty}^{0} \theta v_2^2 \, d\tau - \int_{0}^{u_2(0)} s g(u_1(s), s) \, ds \\ &\geq \int_{-\infty}^{0} \theta v_2^2 \, d\tau - \int_{0}^{u_2(0)} s g(h, s) \, ds \\ &> 0. \end{aligned}$$

The first inequality follows from the fact that $g_1 > 0$, and the integral in the scond term of that line is negative by assumption H6. This chain of inequalities is impossible, so (u, v) cannot be on a connecting orbit. We have thus proved the lemma for the case $u \in \operatorname{int}(N_4)$.

If u is on the left or right edge of N_4 and $v_1 = 0$, then the orbit leaves $\bar{N}_{K,\epsilon}$ in both directions. If $v_2 = 0$ and $v_1 \neq 0$, then the orbit leaves $\bar{N}_{K,\epsilon}$ in one direction. If $u \in N_4 \cap N_5$, i.e., if $u_2 = m$, and $v_2 = 0$, then $\dot{v}_2 > 0$, so the orbit leaves in backward time. If $u_2 = m$, and $v_1 = 0$, then $\mu_1 \dot{v}_1 = -u_1 f(u)$. If $f(u) < 0$, then the orbit leaves $\bar{N}_{K,\epsilon}$ in forward time (since $v_2 > 0$). Similarly, if $f(u) = 0$, then $\ddot{v}_1 = -u_1 v_2 f_2(u) > 0$ so the orbit leaves. If $f(u) > 0$, then $u_1 < a_2$, and if we follow the backward orbit, it moves in the direction of decreasing u_1 (i.e., $v_1 > 0$). This cannot change unless f changes sign (which can only occur of the backward orbit exits N_5 to the left), or u_1 changes sign, which by the remark following the proof of Lemma 3.2 cannot occur on a connecting orbit. Thus (u, v) is not on a connecting orbit if $u_2 = m$ and $v_1 = 0$. The argument for $u \in N_4 \cap N_3$ is similar. □

We now have shown that the only possible internal tangencies can come from connecting orbits which are close to the rest point $M(5)$. The next lemma states that orbits cannot come too close to $M(5)$.

LEMMA 3.4. *There is an $\epsilon > 0$ such that if γ is a connecting orbit then $\gamma \cap B_\epsilon(M(5)) = \emptyset$.*

Proof. Suppose not. Then there is a sequence $\epsilon_n \to 0$ and points (u_n, v_n) on connecting orbits from $M(2)$ to $M(3)$ such that $(u_n, v_n) \in B_{\epsilon_n}(M(5))$. It follows that there is a connecting orbit $(u(t), v(t))$ from $M(2)$ to $M(5)$ (since u_1 is a Lyapunov function). We can again use u_2 as a parameter along the orbit and we obtain

$$\begin{aligned} 0 = \mu_2 \int_{-\infty}^{\infty} v_2 \dot{v}_2 \, d\tau &= \int_{-\infty}^{\infty} \theta v_2^2 \, d\tau - \int_{-\infty}^{0} u_2 v_2 g(u_1, u_2) \, d\tau \\ &= \int_{-\infty}^{\infty} \theta v_2^2 \, d\tau - \int_{0}^{b_5} s g(u_1(s), s) ds \\ &\geq \int_{-\infty}^{\infty} \theta v_2^2 \, d\tau - \int_{0}^{b_5} s g(h, s) ds \\ &> 0. \end{aligned}$$

This is a contradiction. □

Putting together Lemmas 3.1–3.4, it follows that we have an isolating neighborhood for each σ during the deformation, so by Theorem 2.1, there is a $M(2)$ to $M(3)$ travelling wave, and the proof of Theorem 1.2 is complete. □

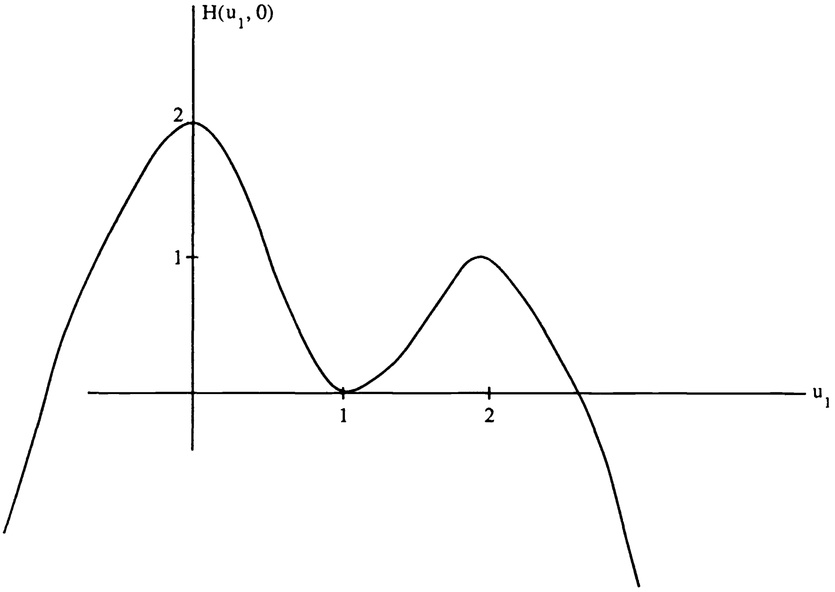


FIG. 4.1

4. The product system. In this section we will analyze the existence of a bistable travelling wave for a very simple system. The purpose of this analysis is to provide a system to which (1.3) can be homotoped in order to prove Theorem 1.3.

Consider the product system

$$(P^\theta) \qquad \begin{aligned} \dot{u}_1 &= v_1 \\ \dot{u}_2 &= v_2 \\ \mu_1 \dot{v}_1 &= \theta v_1 - u_1 h(u_1) \\ \mu_2 \dot{v}_2 &= \theta v_2 - u_2 h(u_2), \end{aligned}$$

where $\theta > 0$ and define $H : \mathbb{R}^2 \to \mathbb{R}$ by

$$H(u_1, u_2) = \int^{u_1} \xi h(\xi)\, d\xi + \int^{u_2} \xi h(\xi)\, d\xi.$$

Furthermore, assume $H(u_1, 0)$ is as in Fig. 4.1, i.e., it has exactly three nondegenerate critical points at $u_1 = 0, 1, 2$ and it takes the values $H(0,0) = 2, H(1,0) = 0$ and $H(2,0) = 1$.

Observe that if we define

$$\mathcal{H}(u, v) = \tfrac{1}{2} \langle v, v \rangle + H(u),$$

then $d\mathcal{H}/d\tau = \theta \langle v, v \rangle$, i.e., $\mathcal{H}$ acts as a Lyapunov function for (P^θ). Fig. 4.2 indicates the (u_1, u_2) coordinates for the set of critical points of (P^θ) and their respective values under $\mathcal{H}$.

Observe that the u_2-axis defines an invariant two-dimensional subspace for the system (P^θ). Let C_3 equal the set of all bounded orbits lying in the space $u_1 = v_1 = 0$. One can check that C_3 is an isolated invariant set.

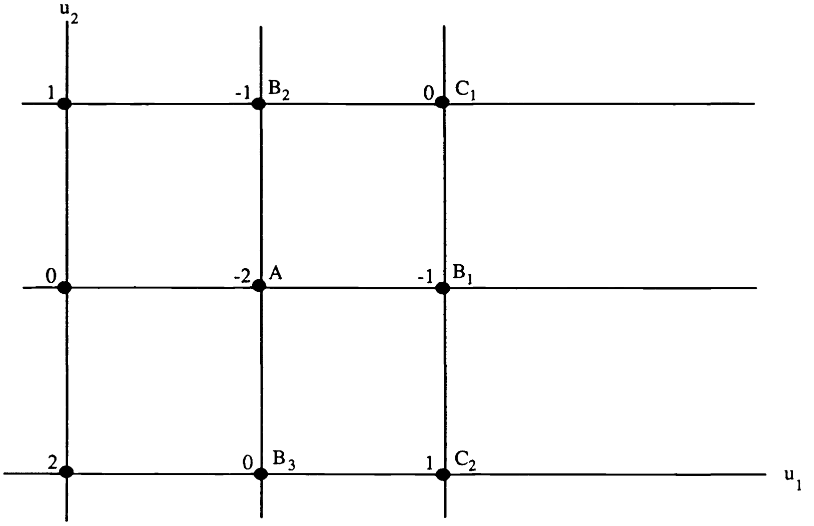


FIG. 4.2

Let S denote the set of bounded solutions to (P^θ), then S is an isolated invariant set and

$$CH_n(S) \approx \begin{cases} \mathbf{Z}_2 & \text{if } n = 2, \\ 0 & \text{otherwise.} \end{cases}$$

A Morse decomposition of S is given by

$$\mathcal{M}(S) = \{A, B_i, C_i \mid i = 1, 2, 3\}.$$

Furthermore, $\mathcal{H}$ and the fact that (P^θ) is a product system implies that the following is an admissible order on $\mathcal{M}(S)$

$$A > B_i > B_3 > C_i > C_3,$$

where $i = 1, 2$. A simple computation leads to the Conley indices of the Morse sets. In particular,

$$CH_n(A) \approx \begin{cases} \mathbf{Z}_2 & \text{if } n = 4, \\ 0 & \text{otherwise,} \end{cases}$$

$$CH_n(B_i) \approx \begin{cases} \mathbf{Z}_2 & \text{if } n = 3, \\ 0 & \text{otherwise,} \end{cases}$$

$$CH_n(C_i) \approx \begin{cases} \mathbf{Z}_2 & \text{if } n = 2, \\ 0 & \text{otherwise.} \end{cases}$$

We now wish to compute the connection matrices of $\mathcal{M}(S)$ for θ small and large. For θ large, the dynamics of (P^θ) are approximated by that of the reaction system (see §8)

$$\dot{u}_i = u_1 h(u_1),$$
$$\dot{u}_2 = u_2 h(u_2).$$

It is left to the reader to check that the appropriate connection matrix in this setting is

$$\Delta^\infty = \begin{array}{c} \\ C_3 \\ C_2 \\ C_1 \\ B_3 \\ B_2 \\ B_1 \\ A \end{array} \begin{array}{c} \begin{array}{ccccccc} C_3 & C_2 & C_1 & B_2 & B_2 & B_1 & A \end{array} \\ \left(\begin{array}{ccccccc} & & & 1 & 1 & 0 & \\ & & & 1 & 0 & 1 & \\ & & & 0 & 1 & 1 & \\ & & & & & & 1 \\ & & & & & & 1 \\ & & & & & & 1 \\ & & & & & & \end{array} \right) \end{array}.$$

For $0 < \theta << 1$, the connection matrix is given by

$$\Delta^0 = \begin{pmatrix} & & & 0 & 0 & 0 & \\ & & & 1 & 0 & 1 & \\ & & & 0 & 1 & 1 & \\ & & & & & & 1 \\ & & & & & & 1 \\ & & & & & & 1 \\ & & & & & & \end{pmatrix}.$$

This is easily obtained by comparing the dynamics of

$$\begin{aligned} \dot{u}_1 &= v_1, \\ \mu_1 \dot{v}_1 &= \theta v_1 - u_1 h(u_1), \end{aligned}$$

with that of

$$\begin{aligned} \dot{u}_1 &= v_1, \\ \mu_1 \dot{v}_1 &= -u_1 h(u_1). \end{aligned}$$

(For more details see Mischaikow [Mi1].)

Since $\mathcal{H}$ acts as a Lyapunov function for all $\theta \in (0, \infty)$ and since the Morse decomposition is valid for all these values of θ, there exists an upper triangular degree-0 isomorphism

$$T = \begin{pmatrix} 1 & x & y & & & & \\ 0 & 1 & z & & & & \\ 0 & 0 & 1 & & & & \\ & & & 1 & a & b & \\ & & & 0 & 1 & 0 & \\ & & & 0 & 0 & 1 & \\ & & & & & & 1 \end{pmatrix},$$

which satisfies the equation (see Franzosa and Mischaikow [F-M])

$$\Delta^0 T + T \Delta^\infty = 0.$$

This implies that $x = y = 1$. Now a result of McCord and Mischaikow [M-M] implies the following.

PROPOSITION 4.1. *The connected simple system for the attractor repeller pair* (C_3, C_1) *differ at* $\theta \approx \infty$ *and* $\theta \approx 0$.

Observe that because (P^θ) is a product system it is easy to check that $S \subset [0,2] \times [0,2] \times \mathbb{R}^2$. Thus, changing the dynamics outside a neighborhood of $[0,2] \times [0,2] \times \mathbb{R}^2$ will have no effect on Proposition 4.1 as long as one restricts attention to the invariant set inside $[0,2] \times [0,2] \times \mathbb{R}^2$. Therefore, we claim that Proposition 4.1 holds for the following system

$$\begin{aligned}\dot{u}_1 &= v_1,\\ \dot{u}_2 &= v_2,\\ \dot{v}_1 &= \theta v_1 - u_1 h_1(u_1,u_2),\\ \dot{v}_2 &= \theta v_2 - u_2 h_2(u_1,u_2),\end{aligned}$$

where $h_i(u_1,u_2) = h(u_i)$ for (u_1,u_2) in a neighborhood of $[0,2] \times [0,2]$ and the zero sets for the functions are as shown in Fig. 4.3.

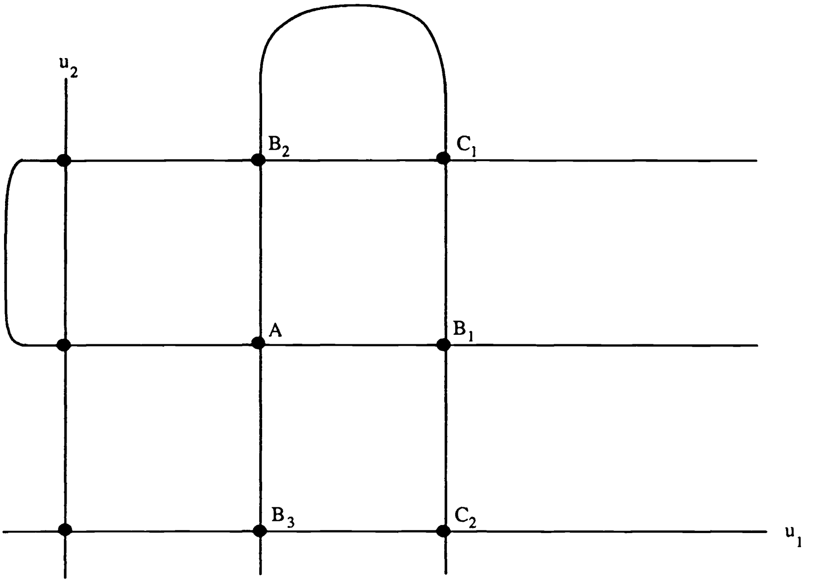


FIG. 4.3

Now recall, that the Conley index is stable under perturbations. Thus Proposition 4.1 holds for the system

$$\begin{aligned}\dot{u}_1 &= v_1,\\ \dot{u}_2 &= v_2,\\ \dot{v}_1 &= \theta v_1 - u_1 f^1(u_1,u_2),\\ \dot{v}_2 &= \theta v_2 - u_2 g^1(u_1,u_2),\end{aligned}$$

where f^1 is a perturbation of h_1 and g^1 is a perturbation of h_2 such that hypothesis H1–H3, H5, H8–H9 are satisfied. The existence of such f^1 and g^1 is clear when one notes that one can choose h such that H5, H8, and H9 are satisfied for system (P^θ).

5. The $M(3) \to M(1)$ travelling wave.

In this section we show that there is a $M(3) \to M(1)$ travelling wave. Choose l, n, d, r, and h as in Fig. 5.1, i.e., the box with corners (d,n), (d,l),(r,n), and (r,l) has its left side in the region $\{f > 0\}$, its right side in the region $\{f < 0\}$, its top in the region $\{g < 0\}$, and its bottom in the region $\{f > 0\}$. Also, choose $r < a_5$, and $n > b_6$. Choose c with $0 < c < a_4$ such that $\int_0^d sg(c,s)ds > 0$, and $\int_c^d sf(s,l)ds > 0$, and choose $h > a_2$.

To prove Theorem 1.3, we use the strengthened version of Theorem 2.1 in which we homotope the system to a product system as discussed in §4. See Fig. 5.2. As usual, we let σ denote the continuation parameter, with $\sigma = 0$ the original system,

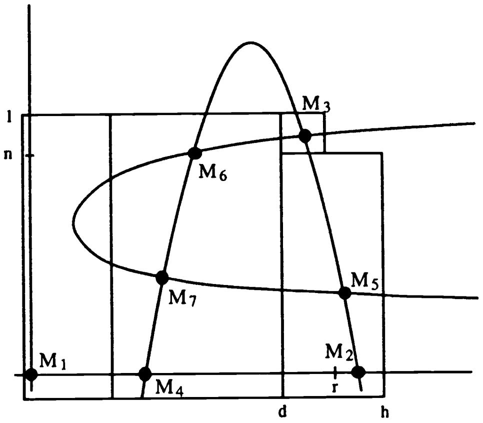


FIG. 5.1

and $\sigma = 1$ the system of Fig. 4.3. Clearly one can choose a deformation so that the assumptions hold for $\sigma \in [0, 1)$. In particular, the integrals in assumptions H8 and H9 can be bounded away from zero, so by the remark following Theorem 1.1, the speeds of any $M(3) \to M(1)$ travelling waves will be positive. For each σ we will construct an isolating neighborhood which contains only the critical points $M(1)$ and $M(3)$, plus any connecting orbits between the points. The construction we make will also construct an isolating neighborhood for $\sigma = 1$, so we can apply Theorem 2.1 and deduce the existence of an $M(3) \to M(1)$ travelling wave.

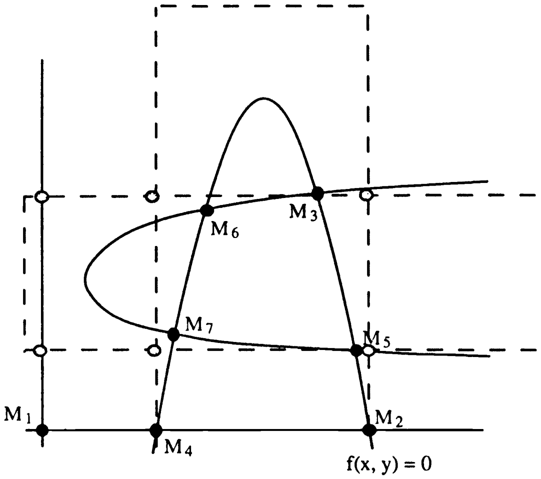


FIG. 5.2

For $e > 0$ define

$$N_1 = [-e, c] \times [-e, l],$$
$$N_2 = [c, d] \times [-e, l],$$
$$N_3 = [d, r] \times [n, l],$$
$$N_4 = [d, h] \times [-e, n].$$

For $K, \epsilon > 0$, define $\bar{N}_1 = N_1 \times [-K, K] \times [-K, K]$, $\bar{N}_2 = N_2 \times [-K, 0] \times [-K, 0]$, $\bar{N}_3 = N_3 \times [-K, K] \times [-K, K]$, $\bar{N}_4 = N_4 \times [-K, K] \times [-K, 0]$, and set

$$\bar{N}_K = \bar{N}_1 \cup \bar{N}_2 \cup \bar{N}_3 \cup \bar{N}_4,$$
$$\bar{N}_{K,\epsilon} = (\bar{N}_1 \cup \bar{N}_2 \cup \bar{N}_3 \cup \bar{N}_4) \setminus (B_\epsilon(2) \cup B_\epsilon(4) \cup B_\epsilon(5) \cup B_\epsilon(6) \cup B_\epsilon(7)).$$

To prove Theorem 1.3, we will show that for K sufficiently large and ϵ sufficiently small, $\bar{N}_{K,\epsilon}$ is an isolating neighborhood for any positive θ. We will use several lemmas in the course of the proof.

LEMMA 5.1. *For any K, $I(\bar{N}_3) = M(3)$.*

Proof. The proof is the same as the proof of Lemma 3.1. □

LEMMA 5.2. *For K sufficiently large, $\bar{N}_1$ is an isolating neighborhood and $I(\bar{N}_1) \subset \{u_1 = v_1 = 0\}$.*

Proof. We first show that every boundary point of $\bar{N}_1$ leaves in forward or backward time. For large K, if $u \in N_1$ and $|v_i| = K$ for $i = 1$ or 2, then the orbit leaves $\bar{N}_1$. If $u_1 = c$ and $v_1 \neq 0$, then the u-projection of the orbit leaves N_1 immediately in forward time if $v_1 > 0$ and in backward time if $v_1 < 0$. If $u_1 = c$ and $v_1 = 0$, then $\mu_1 \dot{v}_1 = -u_1 f(u) > 0$, so the orbit leaves $\bar{N}_1$ in both forward and backward time, i.e., the orbit is externally tangent. A similar argument for the other sides of N_1, i.e., $u_1 = -e$, $u_2 = -e$ and $u_2 = h$, establishes that $\bar{N}_1$ is an isolating neighborhood.

To see that $I(\bar{N}_1) \subset \{u_1 = v_1 = 0\}$, first suppose that $u_1 > 0$ and $v_1 = 0$. Then $\mu_1 \dot{v}_1 = -u_1 f(u) > 0$, i.e., v_1 can only change from negative to positive and in fact, if $v_1 \geq 0$, then v_1 is increasing. So if $u_1 > 0$ and $v_1 \geq 0$, then the u-projection of the forward orbit leaves N_1 to the right (unless it leaves via the top or bottom first). Similarly, if $v_1 < 0$, then v_1 must stay negative in backward time, so the u-projection of the backward orbit leaves N_1. Thus if $u_1 > 0$, the u-projection of the orbit leaves N_1 in one direction or the other. If $u_1 < 0$, a similar argument shows that the orbit exits to the left (in forward time if $v_1 \leq 0$ and in backward time if $v_1 > 0$). Finally if $u_1 = 0$ and $v_1 \neq 0$, then the orbit contains points with nonzero u_1 so the u-projection of the orbit leaves N_1, and it follows that $I(\bar{N}_1) \subset \{u_1 = v_1 = 0\}$. □

The argument used in the second paragraph of this lemma actually shows that if $u_2 < 0$ for any orbit, then $\dot{u}_2$ must stay negative on either the entire forward or entire backward orbit, so any orbit with $u_2 < 0$ somewhere on the orbit cannot be contained in $\bar{N}_{K,\epsilon}$.

Lemmas 5.1 and 5.2, plus the fact that $v_2 \leq 0$ in $\bar{N}_2$ and $\bar{N}_4$, imply that any nonconstant orbit γ which stays in $\bar{N}_{K,\epsilon}$ for all time and is not in the invariant subspace $\{u_1 = v_1 = 0\}$ must have $\omega(\gamma) \subset \{u_1 = v_1 = 0\}$ and $\omega^*(\gamma) = M(3)$. To show that $\bar{N}_{K,\epsilon}$ is an isolating neighborhood, we must show that no connecting orbit in $\bar{N}_{K,\epsilon}$ is internally tangent.

LEMMA 5.3. *Suppose (u, v) is on a connecting orbit in $\bar{N}_K$ (for any K) with $u \in N_2$. Then $v_1 \neq 0$ and $v_2 \neq 0$.*

Proof. To prove this, we will assume $u \in N_2$, $v_i = 0$, and show that (u, v) is not on a connecting orbit. First suppose $u \in \operatorname{int}(N_2)$. If $v_2 = 0$ and $\dot{v}_2 \neq 0$, then the orbit leaves immediately in one direction or the other. If $v_2 = \dot{v}_2 = 0$, then $\mu_2 \ddot{v}_2 = -u_2 v_1 g_1(u)$. If $u_2 \leq 0$ then the orbit cannot be a connection in $\bar{N}_{K,\epsilon}$, so assume $u_2 > 0$. Then $\ddot{v}_2 > 0$, and the orbit leaves immediately in forward time unless $v_1 = 0$. If $v_1 = 0$ and $\dot{v}_1 \neq 0$, then the orbit leaves immediately in one direction or the other, and if $v_1 = \dot{v}_1 = 0$, this combined with $v_2 = \dot{v}_2 = 0$ implies that (u, v) is a critical point. This finishes the case $u \in \operatorname{int}(N_2)$, $v_2 = 0$.

If $u \in \operatorname{int}(N_2)$, $v_1 = 0$ and $\dot{v}_1 = 0$, then $f(u) = 0$ and we may assume $u_2 > 0$. Assume time is parameterized so that $\tau = 0$ at (u, v) and $\tau = t_0 < 0$ when the orbit enters $\bar{N}_2$. If we integrate $\mu_1 v_1 \dot{v}_1$ along the orbit segment in $\bar{N}_2$ we get

$$\begin{aligned} -\mu_1 v_1^2(t_0)/2 = \mu_1 \int_{t_0}^0 v_1\dot{v}_1\,d\tau &= \int_{t_0}^0 \theta v_1^2\,d\tau - \int_{t_0}^0 v_1 u_1 f(u)\,d\tau \\ &= \int_{t_0}^0 \theta v_1^2\,d\tau + \int_{u_1(t_0)}^d u_1 f(u_1, u_2(u_1))\,du_1 \\ &\geq \int_{t_0}^0 \theta v_1^2\,d\tau + \int_{u_1(t_0)}^d sf(s,l)\,ds \\ &> 0. \end{aligned}$$

To get the second line, we parameterize the path by u_1 which we can do since $v_1 < 0$ in $\bar{N}_2$. The inequalities come from the fact that $f_2 > 0$ and H9. The resulting inequality is absurd, so we cannot have $u \in \operatorname{int}(N_2)$, $v_1 = 0$, and $\dot{v}_1 = 0$.

If $u_2 = l$, i.e., u is on the top boundary of N_2, and if $v_2 = 0$, then $\dot{v}_2 > 0$, so the orbit leaves in both directions. If $u_2 = -e$, then the orbit leaves as noted in the remark after the proof of Lemma 5.2. If $u_1 = c$ and $v_1 = 0$, then $\dot{v}_1 < 0$, so the orbit leaves in both directions.

Suppose $u_1 = c$, $u_2 = z$, and $v_2 = 0$. If $\gamma \subset \bar{N}_{K,\epsilon}$, then $v_1 < 0$ as noted above. $\mu_2\dot{v}_2 = -u_2 g(u)$. If $g(u) > 0$, then $\dot{v}_2 < 0$, so the orbit leaves $\bar{N}_{K,\epsilon}$ in backward time. If $g(u) > 0$ and (c,z) lies above the $\{g \geq 0\}$ region, then $\dot{v}_2 > 0$ so v_2 is positive in forward time. In the region $(0,c) \times [z,\infty)$, $g(u) < 0$, so if $v_2 = 0$, then $\dot{v}_2 > 0$, i.e., v_2 cannot change from positive to negative on the forward orbit (unless the orbit crosses $\{u_1 = c\}$ again, in which case it leaves $\bar{N}_{K,\epsilon}$). Since $\omega(\gamma) \subset \{u_1 = v_1 = 0\}$ and the flow on $\{u_1 = v_1 = 0\}$ is gradient-like, $\omega(\gamma)$ must be a critical point. This is impossible since $u_2 \geq z$ on the forward orbit from (c,z). Thus $\gamma \not\subset \bar{N}_{K,\epsilon}$ if γ passes through (c,z) with $v_2 = 0$ for such a z. If z is the larger value in the intersection $\{g = 0\} \cap \{u_1 = c\}$, then $\dot{v}_2 = 0$, but $\ddot{v}_2 = -u_2 v_1 g_1(u) > 0$ so the same argument applies. If $g(c,z) \leq 0$, and (c,z) lies below the $\{g > 0\}$ region, then we integrate $\mu_2 v_2 \dot{v}_2$ along the orbit segment in $\bar{N}_2 \cup \bar{N}_4$. Assume $\tau = 0$ at (c,z), and $\tau = t_0 < 0$ when the orbit enters $\bar{N}_2 \cup \bar{N}_4$. Since $v_2 < 0$ in $\bar{N}_2 \cup \bar{N}_4$, we can parameterize the orbit segment by u_2 and we get

$$\begin{aligned} -\mu_2 v_2^2(t_0)/2 = \mu_2 \int_{t_0}^0 v_2\dot{v}_2\,d\tau &= \int_{t_0}^0 \theta v_2^2\,d\tau - \int_{t_0}^0 v_2 u_2 g(u)\,d\tau \\ &= \int_{t_0}^0 \theta v_2^2\,d\tau + \int_z^{u_2(t_0)} u_2 g(u)\,du_2 \\ &\geq \int_{t_0}^0 \theta v_2^2\,d\tau + \int_z^{u_2(t_0)} sg(c,s)\,ds \\ &\geq \int_{t_0}^0 \theta v_2^2\,d\tau + \int_0^n sg(c,s)\,ds \\ &> 0, \end{aligned}$$

using $g_1 < 0$, the fact that $g(c,y) < 0$ for $y > b_6$, and H9. This can't happen, so v_2 can't be zero when $u_1 = c$ on an orbit lying entirely in $\bar{N}_{K,\epsilon}$. Notice that the same argument using $t = \infty$ instead of $t = 0$ for the upper limit shows that there cannot be an $M(3) \to M(4)$ connection in $\bar{N}_K$.

Finally, if $u_1 = d$ and $v_1 = 0$, $\dot{v}_1 < 0$ so the orbit leaves in both directions, and if $v_2 = 0$, the argument is the same as that for $u_1 = d$. □

LEMMA 5.4. *Suppose (u,v) is on an $M(3) \to M(1)$ connecting orbit in $\bar{N}_K$ (for any K) with $u \in N_4$. Then $v_2 \neq 0$.*

Proof. If $v_2 = 0$ and $v_2 \in \operatorname{int}(N_4)$, then if $\dot{v}_2 \neq 0$, then the orbit leaves $\bar{N}_4$, so assume $v_2 = \dot{v}_2 = 0$. Then either $u_2 = 0$, in whch case (u,v) is in the invariant plane $\{u_2 = v_2 = 0\}$ (and not on a connecting orbit), or $g(u) = 0$. If $g(u) = 0$, then we integrate $\mu_2 v_2 \dot{v}_2$ along the orbit segment in $\bar{N}_4$, setting $\tau = t_0$ as the time when the orbit enters $\bar{N}_4$, and $\tau = 0$ at (u,v). Note that $u_2(t_0) = n$.

$$\begin{aligned}
-\mu_2 v_2^2(t_0)/2 = \mu_2 \int_{t_0}^{0} v_2 \dot{v}_2 \, d\tau &= \int_{t_0}^{0} \theta v_2^2 \, d\tau - \int_{t_0}^{0} v_2 u_2 g(u) \, d\tau \\
&= \int_{t_0}^{0} \theta v_2^2 \, d\tau + \int_{g(0)}^{n} u_2 g(u) \, du_2 \\
&\geq \int_{t_0}^{0} \theta v_2^2 \, d\tau + \int_{g(0)}^{n} s g(c,s) \, ds \\
&\geq \int_{t_0}^{0} \theta v_2^2 \, d\tau + \int_{0}^{n} s g(c,s) \, ds \\
&> 0,
\end{aligned}$$

so we have a contradiction, and there is no such (u,v).

The verification that v_2 cannot be zero on a connecting orbit when $u \in \partial N_4$ is tediously similar to the proof of Lemma 5.3, and hence omitted. □

LEMMA 5.5. *There is no $M(3) \to M(2)$, $M(3) \to M(4)$, $M(3) \to M(5)$, $M(3) \to M(6)$, or $M(3) \to M(7)$ connection in $\bar{N}_K$.*

Proof. Suppose there were an $M(3) \to M(2)$ connection. Then we can parameterize the orbit segment in $\bar{N}_4$ by the u_2 coordinate, and we get the same contradiction as in the proof of Lemma 5.4, so we cannot have a $M(3) \to M(2)$ connection. Similarly, we cannot have a $M(3) \to M(5)$ connection. It was noted in the proof of Lemma 5.3 that there is no $M(3) \to M(4)$ connection. If there were a $M(3) \to M(6)$ connection, parameterize the orbit segment in $\bar{N}_2$ such that $t = 0$ when the orbit enters $\bar{N}_2$, i.e. when $u_1 = d$. Using the fact that u_1 is decreasing along the orbit segment we get

$$\begin{aligned}
-\mu_1 v_1^2(0)/2 = \mu_1 \int_{0}^{\infty} v_1 \dot{v}_1 \, d\tau &= \int_{0}^{\infty} \theta v_1^2 \, d\tau - \int_{0}^{\infty} v_1 u_1 f(u) \, d\tau \\
&= \int_{0}^{\infty} \theta v_1^2 \, d\tau + \int_{a_6}^{d} u_1 f(u) \, du_1 \\
&\geq \int_{-\infty}^{\infty} \theta v_1^2 \, d\tau + \int_{a_4}^{d} s f(s,l) \, ds \\
&> 0,
\end{aligned}$$

with the last inequality following from H8. Thus there is no $M(3) \to M(6)$ connection. Finally, there cannot be a $M(3) \to M(7)$ connection beacuse $M(7)$ is a repeller for $\theta > 0$. □

Proof of Theorem 1.3. The only possible orbits which can stay in $\bar{N}_{K,\epsilon}$ for all time are orbits lying in the invariant plane $\{u_1 = v_1 = 0\}$ and orbits connecting $M(3)$ to a point in the invariant plane. We will show that no such connecting orbit can be internally tangent to the boundary of $\bar{N}_{K,\epsilon}$ for large enough K and small enough ϵ. If K is large enough, then no orbit with $v_i = K$ can stay in $\bar{N}_{K,\epsilon}$ for all time.

It is easy to check that if u lies on the boundary of $N_1 \cup N_2 \cup N_3 \cup N_4$, then using the above arguments the orbit leaves $\bar{N}_{K,\epsilon}$ immediately in one time direction or the other. Lemmas 5.3 and 5.4 show that there cannot be an internal tangency in $\bar{N}_K$ when $v_1 = 0$ or $v_2 = 0$. The only other possile source of internal tangency is around the ϵ balls which are cut out.

Since $M(7)$ is a repeller, we can simply take ϵ such that $\omega^*(B_\epsilon(M(7))) = M(7)$. For the other critical points, suppose that for any $\epsilon > 0$ there is no ϵ-ball around $M(j)$ which can be cut out to give an isolating neighborhood, $j \in \{2,4,5,6\}$. Then there is a sequence $\epsilon_n \to 0$ and a sequence (u^n, v^n), with (u^n, v^n) on an $M(3) \to M(1)$ connection which passes within ϵ_n of $M(j)$. If we parameterize the connecting orbit so that $t = 0$ at (u^n, v^n), then the same integrals which were used to derive a contradiction in the proof of Lemma 4.6 can be used here to derive the same contradiction. Thus no sequence of $\epsilon_n \to 0$ can be found, so for K sufficiently large and ϵ sufficiently small, $\bar{N}_{K,\epsilon}$ is an isolating neighborhood. □

6. Bifurcation results. This section presents the proofs of Theorems 1.4 and 1.5. As will be seen shortly, they follow immediately from work of McCord and Mischaikow [M-M], thus we begin by describing their result.

Let $\Phi : \mathbf{R} \times X \times \Xi \to X$ be a parameter flow (see §2 for a concrete description) where $\Xi = \mathbf{R}^2$, i.e., we are thinking of a system with two independent parameters. Let S be an isolated invariant set under Φ. We need to make several assumptions, the first being the following.

A1. $\mathcal{M}(S) = \{M(p) \mid p = 1,2,3 \text{ and } 3 > 2 > 1\}$ is a Morse decomposition of S. Let $S_\xi = S \cap (X \times \{\xi\})$ and $M_\xi(p) = M(p) \cap (X \times \{\xi\})$. It can be easily checked that S_ξ is an isolated invariant set under φ_ξ and $\mathcal{M}(S_\xi) = \{M_\xi(p) \mid p = 1,2,3,\ 3 > 2 > 1\}$ a corresponding Morse decomposition. We shall also assume the following.

A2. $h(M_\xi(p), \varphi_\xi) \sim \Sigma^n$, $n \geq 1$, $p = 1,2,3$, $\xi \in \Xi$, where Σ^n denotes the pointed n-sphere. Since $\{1,2\}$ and $\{2,3\}$ define intervals in $\{1,2,3\}$ under the ordering on the Morse decomposition,

$$M(i,i+1) = M(i) \cup M(i+1) \cup C(M(i+1),\ M(i))$$

is an isolated invariant set and has an attractor repeller decomposition given by $(M(i),\ M(i+1))$. Let $I_\xi(i,i+1)$ denote the connected simple system associated with $h(M(i,i+1), \varphi_\xi)$. Let $\Xi' \subset \Xi$ such that $M(i,i+1) = M(i) \cup M(i+1)$ *for* $i = 1,2$.

A3. Assume there exist parameter values $\xi_0, \xi_1, \xi_2, \xi_3 \in \Xi'$ for which

$$I_0(1,2), \quad I_1(1,2) \not\cong I_2(1,2), \quad I_3(1,2)$$

and

$$I_0(2,3), \quad I_2(2,3) \not\cong I_1(2,3), \quad I_3(,3);$$

A4.

$$I_0(1,2) \cong I_11,2), \qquad I_2(1,2) \cong I_3(1,2),$$
$$I_0(2,3) \cong I_2(2,3), \qquad I_1(2,3) \cong I_3(2,3).$$

Let $C_{ij} = \{\xi \in \Xi \mid$ there exists a connecting orbit from $M_\lambda(i)$ to $M_\lambda(j)\}$.

THEOREM 6.1 [M-M, Thm. 4.2 and Cor. 4.3]. *Assume* A1–A4, *and assume that* C_{21} *does not separate* ξ_0 *from* ξ_1 *nor* ξ_2 *from* ξ_1 *and that* C_{32} *does not separate* ξ_0 *from* ξ_1 *nor* ξ_2 *from* ξ_3. *Then* $C_{31} \neq \emptyset$. *Furthermore,* $\mathrm{cl}(C_{31}) \cap C_{21} \cap C_{32} \neq \emptyset$.

We shall now show how Theorem 6.1 leads to Theorem 1.4. Hopefully, from the discussion it will become clear why the hypothesis needs to be stated in such a complicated manner.

Proof of Theorem 1.4. Consider Fig. 6.1, which shows $\Xi = [0,1] \times [0,\infty)$. We consider $(\lambda, \theta) = \xi \in \Xi$ where φ_ξ is the flow associated to the ordinary differential equations

$$(6.1^{\lambda,\theta}) \qquad \begin{aligned} \dot{u}_1 &= v_1, \\ \dot{u}_2 &= v_2, \\ \mu_1 \dot{v}_1 &= \theta v_1 - u_1 f^\lambda(u_1, u_2), \\ \mu_2 \dot{v}_2 &= \theta v_2 - u_2 g^\lambda(u_1, u_2). \end{aligned}$$

The results of Theorems 1.1 and 1.2 guarantee that for each fixed $\lambda \in [0,1]$ there exists θ^λ_{12}, θ^λ_{23} wave speeds at which $M(1) \to M(2)$ and $M(2) \to M(3)$ waves exist. Actually, the proof of Theorem 2.1 tells us more. First, there is an upper and lower bound for θ^λ_{ij}. Second, the existence of these wave speeds is obtained by knowledge that the connected simple systems for $h(M(i,j))$ differ at $\theta \approx \infty$ and $\theta \approx 0$. Thus, Figure 6.1 is the simplest possible diagram of the sets θ^λ_{12} and θ^λ_{23}.

Let $\xi_i = (\lambda_i, \theta_i)$. Chose ξ_0 such that θ_0 is greater than θ^λ_{12} and θ^λ_{23} for all $\lambda \in [0,1]$, where θ^λ_{ij} are determined by Theorems 1.1, 1.2, or 1.3. Similarly, ξ_3 is chosen such that $\theta_3 > 0$ is less than $\underline{\theta}^\lambda_{12}$ and $\underline{\theta}^\lambda_{23}$ for all $\lambda \in [0,1]$. Finally, the hypotheses of Theorem 1.4 allows us to choose ξ_1 and ξ_2 such that $\lambda_2 = 0$, $\bar{\theta}^0_{23} < \theta_2 < \underline{\theta}^0_{12}$, $\lambda_1 = 1$, and $\underline{\theta}^1_{23} > \theta_1 > \bar{\theta}^1_{12}$. Because θ^λ_{ij} was determined by changes in the connected simple system of $h(M(i,j))$ one can immediately conclude that

$$\begin{aligned} I_0(1,2) &\cong I_1(1,2), \\ I_2(1,2) &\cong I_3(1,2), \\ I_0(1,2) &\not\cong I_2(1,2), \\ I_1(1,2) &\not\cong I_3(1,2), \end{aligned}$$

and similarly for $I_\xi(2,3)$. Thus, Theorem 6.1 gives the desired result. The proofs for the other i, j follow similarly. □

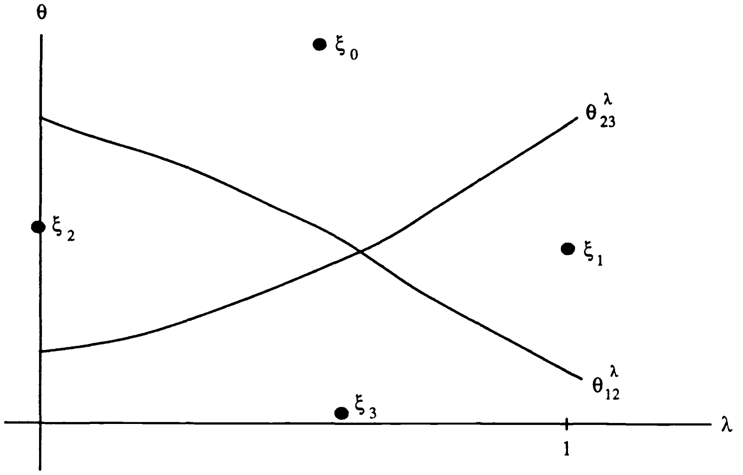


FIG. 6.1

The proof for Theorem 1.5 is similar, if one observes that in the proof of Theorem 6.1 the existence of C_{13} is obtained as the separating set of parameter value for two different connected simple systems for $h(M(13))$.

7. Existence of nonmonotone bistable waves. The results of the previous section were dependent upon knowing that the wave speeds of different bistable waves could be made to cross by changing the nonlinearities. In this section, we intend to show that it is easy to construct parameterized families of nonlinearities for which this phenomenon occurs. One of the implications of our construction is that for a wide class of predator-prey systems of the type being considered here one should expect the existence of nonmonotone travelling waves. It should also be pointed out that the constructions we make are general and applicable to other reaction diffusion systems. To emphasize this latter point we begin with a schematic sketch of the general existence proof (in effect outlining the proof of Theorem 1.4) along with comments of how this will be done for our specific predator-prey system.

Step 1. For a class of reaction terms, one proves the existence of travelling waves $M(i) \rightarrow M(j)$ and $M(j) \rightarrow M(k)$. Furthermore, one obtains the existence results by showing that the connected simple systems of the Conley indices differ; depending on whether the wave speed is large or small.

Comment. This is the content of Theorems 1.1, 1.2, and 1.3 where the reaction terms are assumed to satisfy H1–H9.

Step 2. One parameterizes the class of possible reaction terms. Let Λ denote the set of parameters. Also, let $I^\lambda_{ij} = [\underline{\theta}^\lambda_{ij}, \bar{\theta}^\lambda_{ij}]$ and $I^\lambda_{jk} = [\underline{\theta}^\lambda_{jk}, \bar{\theta}^\lambda_{jk}]$ denote bounded intervals in $\mathbf{R}$ such that if θ^λ_{ij} is a wave speed for which an $M(i) \rightarrow M(j)$ wave occurs, then $\theta^\lambda_{ij} \in I^\lambda_{ij}$.

Comment. In the statement of Theorem 1.4, we chose $\Lambda = [0,1]$ for the sake of simplicity. For more concrete problems there may be a more natural choice for Λ. From a practical point of view the only necessary restrictions on Λ is that it be compact and simply connected.

Furthermore, it is assumed implicitly in the statement of step 2 that the connected simple system of the Conley index for the attractor-repeller pair $(M(j), M(i))$ for wave speeds which lie above I^λ_{ij} differ from the connected simple system for wave speeds which lie below I^λ_{ij}. This will always be the case if the existence of bistable waves were found by an application of Theorem 2.1.

Step 3. One finds parameter values λ_0 and λ_1 such that $\bar{\theta}^{\lambda_0}_{ij} < \underline{\theta}^{\lambda_1}_{ij}$ and $\bar{\theta}^{\lambda_1}_{ij} < \underline{\theta}^{\lambda_0}_{ij}$.

Comment. This of course involves estimates on the wave speeds which may or may not be easy to obtain. In Theorems 7.1, 7.4, and 7.5 we present some general results which allow us to make comparisons of wave speeds between certain nonlinearities. Given these theorems it then becomes a triviality to choose the desired nonlinearities at λ_0 and λ_1.

Step 4. One applies Theorem 6.1.

Comment. See the proof of Theorem 1.4 to see how this is done.

With this outline in mind we now turn to the problem of comparing wave speeds for different nonlinearities. To simplify notation, we consider the one-dimensional problem which has the form

$$\begin{aligned} \dot{u} &= v, \\ \dot{v} &= \theta v - f(u). \end{aligned} \tag{7.1}$$

(Here we have absorbed the diffusion coefficient in θ and f, and we write $f(u)$ instead of $uf(u)$.) The nonlinearity f satisfies $f(0) = f(a) = f(1) = 0$, f is negative between zero and a and positive between a and 1. Let F be such that $F'(u) = f(u)$, $F(0) = 0$, so F has maxima at zero and 1, and a minimum at $u = a$. We assume $F(1) > 0$, so Theorem 1.1 guarantees a positive θ^* so there is bistable wave, i.e., a solution to (7.1) connecting the saddle point (0,0) to the saddle point (1,0). Moreover, it is well known that this connecting orbit satisfies $v > 0$, so we can parameterize the travelling by u instead of t, and (7.1) becomes

$$v\frac{dv}{du} = \theta v - f(u). \tag{7.2}$$

Finally, by integrating (7.2) we get

$$\frac{v^2(u)}{2} = \theta\int_0^u v(\xi)\,d\xi - F(u). \tag{7.3}$$

There are two ways to alter the equation to decrease the wave speed: we can decrease f (near the minimum between zero and a), or we can make the well in F deeper (near $u = a$). Notice that decreasing the well has the effect of decreasing f for $u < a$ and increasing f for $u > a$. We show that either technique can be used to make the wave speed $\theta \to 0$ monotonically.

THEOREM 7.1. *Let $I \subset (0, a)$ be an interval, and let f_2 be a function with wave speed θ_2^*. Then for any $\epsilon > 0$, there is a function f_1 which agrees with f_2 on $[0,1] \setminus I$ and has $0 < \theta_1^* < \epsilon$.*

The proof of this theorem requires two lemmas.

LEMMA 7.2. *Suppose $f_1(u) \leq f_2(u)$ for all u. If $f_1(u) \neq f_2(u)$ for some u, then $\theta_1^* < \theta_2^*$.*

Proof. We have $F_i(u) = \int_0^u f_i(\xi)\,d\xi$ for $i = 1, 2$, and let b denote a point with $f_1(b) < f_2(b)$, so $F_1(u) < F_2(u)$ for $u > b$. Fix $\theta = \theta_2^*$ for both f_1 and f_2. For the f_1 flow, there is a connecting orbit from the $u = 0$ saddle to the $u = 1$ saddle. In the f_2 flow, the corresponding orbit coming from the $u = 0$ saddle will agree with the connecting orbit as long as f_1 and f_2 agree. However, (7.3) implies that for $u > b$, $v_2(u) > v_1(u)$, since $v_1, v_2 \geq 0$, the θ is the same, and $F_1(u) < F_2(u)$. Indeed, (7.3) implies that $v_1 - v_2$ is an increasing function of u, so $v_1(u) > 0$, and $\theta_1^* \neq \theta_2^*$. It is clear from (2) that increasing θ for a fixed f increases $v(u)$, therefore we must have $\theta_1^* < \theta_2^*$. □

As noted in the previous proof, since $F(u) = \int_0^u f(\xi)\,d\xi$, decreasing f causes a corresponding decrease in F, so by decreasing f in some interval we can force $F(1) - F(0)$ to become arbitrarily small. (If the difference becomes nonpositive, θ^* will also become nonpositive.) Since $\theta\int_0^1 v\,du = F(1) - F(0)$, we only need to show that $\int_0^1 v\,du$ is bounded away from zero as f is decreased.

LEMMA 7.3. *Suppose $f_1(u) \leq f_2(u)$ for all u. Then $\int_0^1 v_1\,du \geq \int_0^a \sqrt{-2F_2(\xi)}\,d\xi$.*

Proof. By (7.3), $v_1^2(u)/2 > -F_1(u)$ and by hypothesis $F_1(u) \leq F_2(u)$. For $0 < u < a, F_i(u) < 0$, so we have

$$\begin{aligned}\int_0^1 v_1\,du &\geq \int_0^a v_1\,du \\ &\geq \int_0^a \sqrt{-2F_1(u)}\,du \\ &\geq \int_0^a \sqrt{-2F_2(u)}\,du. \qquad \square\end{aligned}$$

Proof of Theorem 7.1. Theorem 7.1 is an easy consequence of Lemmas 7.2 and 7.3. Start with some f_2, and let $k = \int_0^a \sqrt{-2F_2(u)}\,du$. Now decrease the value of f in

I (this introduces no new critical points), so $F(1) - F(0) < \epsilon/k$. For this f we have $\theta\int_0^1 v_1\,du = F(1) - F(0) < \epsilon/k$ and $\int_0^1 v_1\,du \geq k$ by Lemma 3, so $\theta < \epsilon$. □

The arguments to show that increasing the depth of the well of F can force the wave speed to go to zero are similar. We now state assumptions in terms of F.

THEOREM 7.4. *Let F_2 be a function with wave speed θ_2^*, and let I be an interval containing a with $F_2 < 0$ on I. Then for any $\epsilon > 0$, there is a function F_1 which agrees with F_2 on $[0,1] \setminus I$ and has $0 < \theta_1^* < \epsilon$.*

We use one more lemma.

LEMMA 7.5. *Suppose $F_1(u) \leq F_2(u)$ for all u. If $F_1(u) \neq F_2(u)$ for some u, then $\theta_1^* < \theta_2^*$.*

Proof. The proof is the same as the proof of Lemma 2. □

Proof of Theorem 7.4. Start with F_2, and let $F_2(1) - F_2(0) = k$. Now decrease F in I so that $\int_0^a \sqrt{-2F(u)}\,du > k/\epsilon$. Since $a \in I$, this can be done without introducing new critical points, and since $I \subset (0,1)$ the new F still satisfies $F(1) - F(0) = k$. We have

$$\begin{aligned} k = F(1) - F(0) &= \theta^* \int_0^1 v\,du \\ &\geq \theta^* \int_0^a \sqrt{-2F(u)}\,du \\ &> \theta^* k/\epsilon, \end{aligned}$$

so $\theta^* < \epsilon$. □

We now show that by raising $F(1)$ (or equivalently, lowering $F(0)$), we can force θ^* to go to ∞.

THEOREM 7.6. *Let F_2 be a function with wave speed θ_2^*, and let $b \in (a,1)$ be the point with $F_2(b) = 0$. Then for any $M, \epsilon > 0$, there is a function F_1 which agrees with F_2 on $(b+\epsilon, 1]$ and has $\theta_1^* > M$.*

Proof. Let v be a solution to (7.2) with $\theta = M$, i.e., v satisfies $dv/du = M - f_2(u)/v$, $v(0) = 0$ where $F_2' = f_2$. Let $\gamma = \max\{v(u) \mid 0 \leq u \leq b\}$. Note that $f_2(u) > 0$ for $u \in [b,1]$, so $dv/du \leq M$ for $u \in [b,1]$. It follows that $v(u) \leq \gamma + M(1-b)$ for all u. Make a function $F_3(u)$ which agrees with $F_2(u)$ for $u \in [0,1]$, $F_3' > 0$ in $[b,1]$, (so the bound on $v(u)$ holds for solutions involving F_3 and $\theta = M$), and $F_3(u) = F_2(u) + K$ for $u \in [b+\epsilon, 1]$ where K is a constant to be named later. Equation (7.3) implies $F_3(1) = M\int_0^1 v(\xi)\,d\xi$ if $v > 0$ along the whole orbit. But $M\int_0^1 v(\xi)\,d\xi < M(\gamma + M(1-b)) < F_3(1)$, if K is large enough, so the wave speed for F_3 is greater than M. We now set $F_1(u) = F_3(u) - K$. □

Obviously, these results hold if the critcal points are at points other that zero, a, and 1.

We now show how these one-dimensional results can be applied to the predator-prey systems we have been studying. We will show that there are systems of the form (1.3) which have $M(1) \to M(3)$ travelling waves and homoclinic $M(1) \to M(1)$ travelling waves.

THEOREM 7.7. *There are systems of the form* (1.3) *satisfying* H1–H3 *and* H5–H7 *which have an $M(1) \to M(3)$ travelling wave with $\theta_{13} > 0$.*

Proof. Theorem 1.2 guarantees an $M(2) \to M(3)$ travelling wave of speed $\theta_{23} > 0$. Let $\underline{\theta}_{23}$ be the minimum speed for any $M(2) \to M(3)$ travelling wave. It is easy to see that the set of wave speeds is closed, so $\underline{\theta}_{23} > 0$. Similarly, let $\bar{\theta}_{23}$ be the largest $M(2) \to M(3)$ wave speed. It will be shown in §9 that this is finite. The idea of the proof is to deform f away from this wave to change the speed of the $M(1) \to M(2)$

travelling wave. So start with any system (1.3) satisfying H1–H3, H5–H7. We apply Theorem 7.6 to the one-dimensional problem

$$\begin{aligned} \dot{u}_1 &= v_1 \\ \dot{v}_1 &= \mu_1 \theta v_1 - u_1 f(u_1, 0) \end{aligned}$$

which is the equation governing the $M(1) \to M(2)$ travelling wave. By altering $f(u_1, 0)$ on the segment $\{ (u_1, 0) \mid \epsilon \leq u_1 < b + \epsilon \}$ (where the b is as in Theorem 7.6, and ϵ is small), we can produce a function $\tilde{f}(u_1, 0)$ so that θ_{12}, the speed of the $M(1) \to M(2)$ travelling wave, is greater than $\bar{\theta}_{23}$. Let U be a neighborhood of the segment $\{ (u_1, 0) \mid \epsilon \leq u_1 < b + \epsilon \}$ which misses the (u_1, u_2)-projection of any $M(2) \to M(3)$ travelling wave. Define f^0 to be a function which satisfies H1–H3, H5–H7, $f^0 = f$ outside of U and $f^0(u_1, 0) = \tilde{f}(u_1, 0)$. Similarly, by using Theorem 7.1 we can alter $f(u_1, 0)$ on the segment $\{ (u_1, 0) | \epsilon \leq u_1 < b + \epsilon \}$ to $\hat{f}(u_1, 0)$ so that θ_{12} is less than $\underline{\theta}_{23}$. Let f^1 be a function which satisfies H1–H3, H5–H7, $f^1 = f$ outside of U and $f^1(u_1, 0) = \hat{f}(u_1, 0)$. Finally, for $0 \leq \lambda \leq 1$ define $f^\lambda(u) = (1 - \lambda) f^0(u) + \lambda f^1(u)$.

Now consider the parameterized family

$$\begin{aligned} \dot{u}_1 &= v_1 \\ \dot{u}_2 &= v_2 \\ \mu_2 \dot{v}_2 &= \theta v_2 - u_2 g(u_1, u_2)\,. \end{aligned}$$

For each $\lambda \in [0, 1]$ there is an $M(1) \to M(2)$ travelling wave of speed θ^λ_{12} and an $M(2) \to M(3)$ travelling wave of speed θ_{32}. By the construction of f^λ, $\theta^0_{12} > \bar{\theta}_{32}$, and $\theta^1_{12} < \underline{\theta}_{32}$. Theorem 1.4 implies the existence of a λ so that the system with f^λ has an $M(1) \to M(3)$ travelling wave of speed $\theta_{13} > 0$. □

We remark that the proof of Theorems 7.7 and 1.4 actually imply that the $M(1) \to M(3)$ travelling wave passes close to $M(2)$, and that θ_{13} is close to the speed of some $M(2) \to M(3)$ travelling wave.

Finally, we discuss how to use Theorem 1.5 to construct homoclinic $M(1) \to M(1)$ travelling waves. We first use Theorem 1.4 to construct $M(2) \to M(1)$ travelling waves (via $M(3)$), then apply Theorem 1.5 to the $M(2) \to M(1)$ and $M(1) \to M(2)$ waves, where we control the speed of the $M(1) \to M(2)$ wave as in the proof of Theorem 7.7. We begin by modifying f so that the critical points M_2, M_5, and M_3 all lie in the same vertical line in the (u_1, u_2)-plane, i.e., $a_2 = a_5 = a_3$. See Fig. 7.1.

The proof of Theorem 1.3 implies that there is an $M(3) \to M(1)$ travelling wave with positive wave speed θ_{31}. Let $\underline{\theta}_{31}$ be the minimum $M(3) \to M(1)$ speed, and $\bar{\theta}_{31}$ be the maximum speed. The one-dimensional analysis applies in the set $\{ u_1 = a_2, v_1 = 0 \}$ to allow us to control the speed θ_{23} of the $M(2) \to M(3)$ travelling wave. As in the proof of Theorem 7.7, we can alter g in a neighborhood of the $u_1 = a_2$ line (in particular, away from any $M(3) \to M(1)$ travelling wave) to obtain functions g^0 so that the speed of the $M(3) \to M(2)$ travelling wave for the system using g^0 is $\theta^0_{23} > \bar{\theta}_{31}$. Similarly, we can alter g to get g^1 so the speed of the $M(3) \to M(2)$ travelling wave for the system using g^1 is $\theta^1_{23} < \underline{\theta}_{31}$. Notice that g^0 and g^1 can be constructed so that hypotheses H6, H7, and H9 are still satisfied. We define $g^\lambda(u) = (1 - \lambda) g^0(u) + \lambda g^1(u)$, and consider the parameterized family

$$\begin{aligned} \dot{u}_1 &= v_1, \\ \dot{u}_2 &= v_2, \\ \mu_1 \dot{v}_1 &= \theta v_1 - u_1 f(u_1, u_2), \\ \mu_2 \dot{v}_2 &= \theta v_2 - u_2 g^\lambda(u_1, u_2). \end{aligned} \tag{7.4}$$

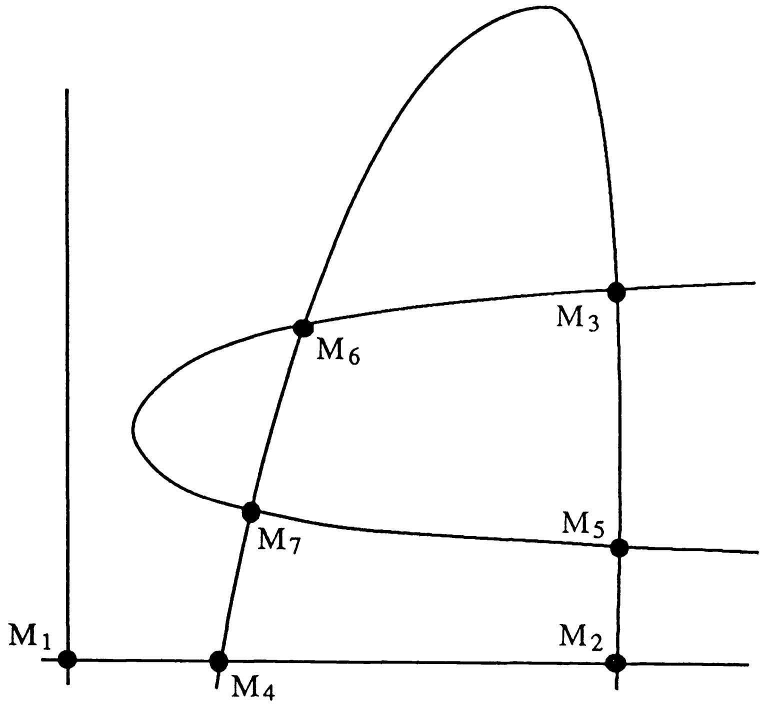


FIG. 7.1

We now argue as in the proof of Theorem 7.7 and obtain a g^λ such that there is an $M(2) \to M(1)$ travelling wave of speed $\theta_{21} > 0$. Note that this wave passes close to $M(3)$.

Now we perturb f slightly so that the points M_2, M_5, and M_3 are no longer in the same line in the (u_1, u_2)-plane, but instead that f satisfies H1–H9. As noted in the proof of Theorem 1.4, because the connected simple systems continue, for a small perturbation the speed of any $M(3) \to M(2)$ travelling wave in the perturbed system will be close to the speed in the system with the vertical line analyzed in the preceeding paragraph. It follows that we can still construct g^0 and g^1 so that $\underline{\theta}^0_{23} > \bar{\theta}_{31}$ and $\bar{\theta}^1_{23} < \underline{\theta}_{31}$, so for the perturbed f there is a g^λ so that system (7.4) has an $M(2) \to M(1)$ travelling wave of speed $\theta_{21} > 0$. We have proved the following.

THEOREM 7.8. *There are systems of the form* (1.3) *satisfying* H1–H9 *which have an* $M(2) \to M(1)$ *travelling wave with* $\theta_{21} > 0$.

Now to show the existence of an $M(1) \to M(1)$ travelling wave, we start with a system from Theorem 7.8, with a $M(2) \to M(1)$ travelling wave of speed $\theta_{21} > 0$. We now repeat the proof of Theorem 7.7 with θ_{21} replacing θ_{32}. Specifically, by changing f near the u_1 axis in the (u_1, u_2)-plane (in particular, away from the $M(2) \to M(1)$ travelling wave), we construct a one-parameter family f^λ so that the system (7.4) satisfies the following.

(i) There is an $M(2) \to M(1)$ travelling wave of speed θ_{21} for all λ (the wave speed is independent of λ since f^λ is unchanged near the wave).

(ii) When $\lambda = 0$, there is an $M(1) \to M(2)$ travelling wave of speed $\theta^0_{12} > \bar{\theta}_{21}$.

(iii) When $\lambda = 1$, there is an $M(1) \to M(2)$ travelling wave of speed $\theta^1_{12} < \underline{\theta}_{21}$.

The conditions of Theorem 1.5 are satisfied, so there is a $\lambda \in (0,1)$ and a homoclinic $M(1) \to M(1)$ travelling wave of speed $\theta^\lambda_{11} > 0$. So we have proved the following.

THEOREM 7.9. *There are systems with $M(1) \to M(1)$ homoclinic travelling waves of speed $\theta_{11} > 0$.*

8. Fisher waves and higher-dimensional analogues. Proving the existence of bistable waves as described in the previous sections was dependent upon careful selections of isolating neighborhoods and homotopies to simpler systems. As will be shown now, for high wave speeds the existence of Fisher waves and their higher-dimensional analogues can be determined directly from the dynamics of the reaction system. In particular we shall state an abstract existence theorem and a simple corollary, followed by a discussion of how they can be applied to the two species predator-prey systems being considered in this paper.

The abstract existence theorem relates, in a general setting, connecting orbits for the reaction system with connecting orbits (i.e., travelling waves) for the travelling wave system. So consider the following n-dimensional system of ordinary differential equations

$$\dot{u} = F(u) \tag{8.1}$$

and the corresponding $2n$-dimensional travelling wave system

$$\begin{aligned} \dot{u} &= v, \\ D\dot{v} &= \theta v - F(u), \end{aligned} \tag{8.2}$$

where D is a diagonal matrix with strictly positive entries and $(u,v) \in \mathbf{R}^n \times \mathbf{R}^n$. For the remainder of this section N will always denote an isolating neighborhood for (8.1) and $S = I(N)$, the maximal invariant set in N. Let ϕ^θ denote the flow on $\mathbf{R}^{2n}$ generated by (8.2). Let $\mathcal{M}(S) = \{M(p) \mid p \in (P,>)\}$ be a Morse decomposition for S and let $\mathcal{CM}(S)$ denote the set of connection matrices for $\mathcal{M}(S)$.

THEOREM 8.1. (i) *There exist positive constants K, Θ such that $N \times [-K,K] \subset \mathbf{R}^n \times \mathbf{R}^n$ is an isolating neighborhood for the flow ϕ^θ when $\theta > \Theta$.*

(ii) *Let $S^\theta = I(N \times [-K,K], \phi^\theta)$. Then, there exists a Morse decomposition of the form*

$$\mathcal{M}(S^\theta) = \{M(p^\theta) \mid p^\theta \in (P,>)\}$$

such that $M(p^\theta)$ and $M(p^{\bar{\theta}})$ are related by continuation for all $\theta, \bar{\theta} > \Theta$.

(iii) *If $\Delta^\theta \in \mathcal{CM}(S^\theta)$, the set of connection matrices for $\mathcal{M}(S^\theta)$, then Δ^θ is a degree n conjugation of Δ for some $\Delta \in \mathcal{CM}(S)$.*

Theorem 8.1 is considerably more general that what we need. For our purposes the following corollary will suffice.

COROLLARY 8.2. *Assume that:*

(i) *For all $p \in P$, $M(p)$ is a hyperbolic fixed point of* (8.1),

(ii) *$\mathcal{CM}(S)$ consists of a unique connection matrix Δ,*

(iii) *$p, q \in P$ are adjacent and $\Delta(p,q) \neq 0$.*

Then there exists Θ such that for $\theta > \Theta$:

(a) *$\mathcal{M}(S) = \{(M(p), 0) \mid p \in (P,>)\}$ is a Morse decomposition for S^θ,*

(b) *Δ^θ, a degree n conjugation of Δ, is the unique connection matrix for $\mathcal{M}(S^\theta)$,*

(c) *there exists a solution* $(u(t), v(t))$ *of* (8.2) *such that*

$$\lim_{t\to\infty}(u(t), v(t)) = (M(q), 0),$$
$$\lim_{t\to-\infty}(u(t), v(t)) = (M(p), 0).$$

The proofs of Theorem 8.1 and Corollary 8.2 can be found in §9. For the moment we concentrate on showing how Corollary 8.2 can be applied to two-species predator-prey systems. In particular, both Theorem 8.1 and Corollary 8.2 suggest that we begin our analysis by investigating the possible connection matrices for the predator-prey system. This has been done by Reineck [Re2] under hypothesis H1, H2, and H3. In particular, $\mathcal{M}(S) = \{M(p) = M_p \mid p = 1, 2, \ldots, 7\}$ forms a Morse decomposition of S, the set of bounded solutions in the positive orthant for the reaction system (8.1). The connection matrix can now be consider as map on

$$CH_0(M_1) \oplus CH_0(M_2) \oplus CH_0(M_3) \oplus CH_1(M_4) \oplus CH_1(M_5) \oplus CH_1(M_6) \oplus CH_2(M_7).$$

Thus it is a 7×7 matrix and since it is degree -1 it must be of the form

$$\Delta = \begin{array}{c|ccccccc} & M_1 & M_2 & M_3 & M_4 & M_5 & M_6 & M_7 \\ \hline M_1 & 0 & 0 & 0 & * & * & * & 0 \\ M_2 & 0 & 0 & 0 & * & * & * & 0 \\ M_3 & 0 & 0 & 0 & * & * & * & 0 \\ M_4 & 0 & 0 & 0 & 0 & 0 & 0 & * \\ M_5 & 0 & 0 & 0 & 0 & 0 & 0 & * \\ M_6 & 0 & 0 & 0 & 0 & 0 & 0 & * \\ M_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array},$$

where $*$ denotes entries which need to be determined. The following theorem provides this information.

THEOREM 8.3 [R2, §3]. *Under the hypothesis* H1, H2, *and* H3, *only the following* 6 *connection matrices can be realized.*

$$B = \begin{array}{c|ccccccc} & M_1 & M_2 & M_3 & M_4 & M_5 & M_6 & M_7 \\ \hline M_1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ M_2 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ M_3 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ M_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ M_5 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ M_6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ M_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array},$$

$$C = \begin{array}{c|ccccccc} & M_1 & M_2 & M_3 & M_4 & M_5 & M_6 & M_7 \\ \hline M_1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ M_2 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ M_3 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ M_4 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ M_5 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ M_6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ M_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array},$$

$$D = \begin{array}{c|ccccccc|} & M_1 & M_2 & M_3 & M_4 & M_5 & M_6 & M_7 \\ M_1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ M_2 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ M_3 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ M_4 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ M_5 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ M_6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ M_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array},$$

$$E = \begin{array}{c|ccccccc|} & M_1 & M_2 & M_3 & M_4 & M_5 & M_6 & M_7 \\ M_1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ M_2 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ M_3 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ M_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ M_5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ M_6 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ M_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array},$$

$$F = \begin{array}{c|ccccccc|} & M_1 & M_2 & M_3 & M_4 & M_5 & M_6 & M_7 \\ M_1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ M_2 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ M_3 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ M_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ M_5 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ M_6 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ M_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array},$$

$$G = \begin{array}{c|ccccccc|} & M_1 & M_2 & M_3 & M_4 & M_5 & M_6 & M_7 \\ M_1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ M_2 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ M_3 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ M_4 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ M_5 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ M_6 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ M_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}.$$

In order to simplify matters we shall make the following assumption.

H10. *The reaction system* (8.1) *is Morse–Smale.*

By Reineck [Re1], H10 implies that the connection matrix is unique. Furthermore, if any entry in the matrix is nonzero then the Morse sets are adjacent, and therefore Corollary 8.2 is applicable. This raises the question, which of the six possible connection matrices correspond to the systems we have been studying. Before providing an answer let us introduce a formal hypothesis. Recall from H2 that $u_2 = \xi_2$ is the point at which $q(u_2)$, the zero set of g, reaches an absolute minimum. Also recall that $M_4 = (a_4, b_4)$.

H11. $q(\xi_2) \leq a_4$

Implicit in the discussion of [Re2] is the following result.

PROPOSITION 8.4. *Assume* H1, H2, H3, *and* H10. *If in addition one assumes*:

(a) H5, *then* $\Delta = E, F,$*or* G;

(b) H11, *then* $\Delta = C, D,$ *or* G;

(c) H5 *and* H10, *then* $\Delta = G$.

Recall that H5 was essential to our construction of an isolating neighborhood for the $M(2) \to M(3)$ bistable wave. From Fig. 5.1 we see that we are assuming H11 in our proof of the existence of the $M(3) \to M(1)$ wave. Therefore, by Proposition 8.4 for hypotheses under which the results of §6 hold, the connection matrix for the reaction system is G. Therefore, as an example of how to apply Propositions 8.2 and 8.4 we state the following result.

PROPOSITION 8.5. *Given* H1–H3, H5, H10, *and* H11, *there exists* $\Theta > 0$ *such that for all* $\theta > \Theta$ *there exist Fisher waves of the form*

$$\begin{aligned} &M(4) \to M(1), &&M(4) \to M(2),\\ &M(5) \to M(2), &&M(5) \to M(3),\\ &M(6) \to M(1), &&M(6) \to M(3). \end{aligned}$$

While Corollary 8.2 is easy to state and easy to apply, its hypotheses are somewhat stronger that desired. In particular, even if 8.2(iii) is violated, i.e., $\Delta(p,q) = 0$, it is possible to have an $M(p) \to M(q)$ travelling wave. As an example consider the following result.

PROPOSITION 8.6. *Assume* H1–H3, H5, H10, *and that* $\Delta = E$. *Then, there exists* $\Theta > 0$ *such that for all* $\theta > \Theta$ *there exists Fisher waves of the form*

$$\begin{aligned} &M(4) \to M(1), &&M(4) \to M(2),\\ &M(5) \to M(2), &&M(5) \to M(3), \end{aligned}$$

and two distinct waves

$$M(6) \to M(3).$$

Proof. For the reaction system (1.2) one can always find an isolating neighborhood which contains an $M_6 \to M_3$ connecting orbit (see [Re2, §3]) and for which, restricted to that neighborhood, the connection matrix entry $\Delta(6,3) = 1$. Since, for $\Delta = E$, $\Delta(6,3) = 0$, there exists another connecting orbit (this follows directly from [McC, Thm. 2.5]). Now apply Theorem 8.1. □

This result should be contrasted with those of one-dimensional systems where the Fisher waves have been most extensively considered. First, for a given wave speed in one species systems there can be at most one Fisher wave connecting two stationary solutions. For two species problems this is no longer the case. Second, in the one species case the minimum wave speed which these Fisher waves occur is associated with the appearance of a bistable wave. This wave speed is known to be unique for the one species problem. In the two species model there is no reason to assume that the minimum wave speed agrees for the two distinct Fisher waves.

We now turn to the question of determining the existence of HDF waves. Recall that for the systems being consider in this paper these take the form $M(7) \to M(j)$ where $j = 1,2,3$. The key to proving the existence of these wave is the following theorem.

THEOREM 8.7 [M-R, Thm. 3.2]. *Let* Δ^θ *be the unique connection matrix for the Morse decomposition* $\mathcal{M}(S^\theta) = \{M(p^\theta) \mid p^\theta \in (P, >^\theta)\}$ *where* $>^\theta$ *is the flow defined order. Let* $\{\alpha, \beta^\pm, \gamma\} \subset P$. *Assume*:

(i) $\gamma >^\theta \beta^\pm >^\theta \alpha$,

(ii) $\{\alpha, \beta^+, \gamma\}$ *and* $\{\alpha, \beta^-, \gamma\}$ *are intervals in* P,

(iii) $\Delta(\beta^\pm, \alpha) \circ \Delta(\gamma, \beta^\pm) \neq 0$,

(iv) $C\check{H}_*(M(p); \mathbf{Z}_2)$ *is finitely generated for* $p \in \{\alpha, \beta^\pm, \gamma\}$.

Then there exists a connected set $\Gamma \subset C(M(\gamma), M(\alpha))$ such that $\mathrm{cl}(\Gamma) \cap M(\beta^{\pm}) \neq \emptyset$.

Obviously, given H1–H3 and H10, Theorems 8.1 and 8.3 determine the possibilities for Δ^{θ}. Now to apply Theorem 8.7 we need only find four element subsets of $\{1,2,3,4,5,6,7\}$ for which the hypotheses apply. As an example of this assume $\Delta^{\theta} = G$. Consider the sets

$$\{\alpha, \beta^{+}, \beta^{-}, \gamma\} = \{7,6,4,1\}, \quad \{7,5,4,2\}, \quad \text{or } \{7,6,5,3\}.$$

It is easy to check that they satisfy hypotheses (i)–(iv), and hence, Theorem 8.7 provides the existence of HDF waves. In fact, since we know that $M(7)$ is a hyperbolic critical point with a four-dimensional unstable manifold and $M(i)$, $i = 1,2,3$, are hyperbolic critical points with two-dimensional unstable manifolds we can conclude that the set of connecting orbits $M(7) \to M(i)$, $i = 1,2,3$, contains open 2-disks. For a more detailed discussion of Theorem 8.7, possible variations, and its applications we refer the reader to [M-R].

Aside from demonstrating how to find Fisher waves, this section serves another purpose, and that is to show how far we are from a general understanding of the existence of travelling waves for predator-prey systems. As was mentioned before, to obtain the types of results presented in §6, it must be that the connection matrix for the reaction system is of the form G. In particular, if the connection matrix for the reaction system is of the form B or D (and Theorem 8.3 guarantees the existence of such systems), then our analysis for the existence of either an $M(2) \to M(3)$ or $M(3) \to M(1)$ travelling wave fails. It is easy to see why this is the case; all our constructions of isolating neighborhoods in §§3 and 5 were based on finding monotone travelling waves. However, the obvious homotopies for the latter two cases suggest the existence of a nonmonotone wave. It seems reasonable to speculate that such waves could be of physical interest if one assumes that the diffusion coefficients are small while the reaction terms are large. If this is indeed the case, then this suggests an important future avenue of research.

9. Proof of Theorem 8.1.

Although it will be clear from the proof, we wish to emphasize that Theorem 8.1 is essentially a corollary of perturbation results due to Conley and Fife [C-F].

Proof of Theorem 8.1. Consider the system of equations

$$(9.1)\qquad \begin{aligned} u' &= \epsilon F(u) + \epsilon w, \\ w' &= D^{-1}w - \epsilon \nabla F(u) \cdot Dw - \epsilon \nabla F(u) \cdot F(u). \end{aligned}$$

For $\epsilon = 0$, (9.1) becomes

$$(9.2)\qquad u' = 0, \qquad w' = D^{-1}w.$$

Since D^{-1} is a diagonal matrix with strictly positive entries, $\{(u,w) \mid w = 0\}$ defines a critical manifold [C-F, p. 160]. The corresponding limit equation [C-F, p. 164] is

$$(9.3)\qquad u' = F(u).$$

If one sets $\epsilon = c^{-2}$, $' = 1/c$ $\dot{}$, and $Dw = \theta v - F(u)$, then (9.1) becomes (8.2), the travelling wave system. One now applies Theorems 3.5A and 3.5B of [C-F] to obtain (i), (ii), and the fact that for every attractor repeller pair of $\mathcal{M}(S^{\theta})$ the connecting maps are n-fold suspensions of the connecting maps for the corresponding attractor-repeller pair in $\mathcal{M}(S)$. (iii) now follows from [Fr2, Thm. 5.7]. □

Proof of Corollary 8.2. An easy eigenvalue computation or [C-F] guarantees that if M_p is a hyperbolic fixed point for (8.1), then so is $(M_p, 0)$ for (8.2) and θ large. Thus (a) follows from Theorems 3.5A and 3.5B of [C-F]. Observe that (b) follows from 8.2(ii) and 8.1(iii), while (c) follows from 8.2(iii) and properties of the connection matrix. □

Acknowledgments. We would like to thank J. Dockery and Y. Hosono for valuable comments and suggestions. We would also like to thank the Institute for Mathematics and its Applications for support and for a stimulating environment in which to do mathematics.

REFERENCES

[Co] C. C. Conley, *Isolated Invariant Sets and the Morse Index*, CBMS Lecture Notes 38, American Mathematical Society, Providence, RI, 1978.

[C-F] C. C. Conley and P. Fife, *Critical manifolds, travelling waves and an example from population genetics*, J. Math. Biol., 14 (1982), pp. 159–176.

[C-G] C. C. Conley and R. Gardner, *An application of the generalized Morse index to travelling wave solutions of a competitive reaction diffusion model*, Indiana Univ. Math. J., 33 (1989), pp. 319–343.

[F-T] M. Feinberg and D. Terman, *Travelling composition waves on isothermal catalyst surfaces*, preprint.

[Fi] P. C. Fife, *Mathematical Aspects of Reacting and Diffusing Systems*, Lecture Notes in Biomath. 28, Springer-Verlag, New York, 1979.

[Fr] R. Franzosa, *The connection matrix theory for Morse decompositions*, Trans. Amer. Math. Soc., 311 (1989), pp. 561–592.

[Fr2] ———, *The continuation theory for connection matrices and Morse decompositions*, Trans. Amer. Math. Soc., 310 (1988), pp. 781–803.

[F-M] R. Franzosa and K. Mischaikow, *Algebraic Transition Matrices*, in preparation.

[Ga] R. Gardner, *Existence of travelling wave solution of predator-prey systems via the Conley index*, SIAM J. Appl. Math., 44 (1984), pp. 56–76.

[G-J] R. Gardner and C. Jones, *Stability of travelling wave solutions of diffusive predator-prey systems*, Trans. Amer. Math. Soc., to appear.

[G-S] R. Gardner and J. Smoller, *The existence of periodic travelling waves for singularly perturbed predator-prey equations via the Conley index*, J. Diff. Eqns., 47 (1983), pp. 133–161.

[McC] C. McCord, *The connection map for attractor-repeller pairs*, Trans. Amer. Math. Soc., 307 (1988), pp. 195–203.

[M-M] C. McCord and K. Mischaikow, *Connected Simple Systems, Transition Matrices, and Heteroclinic Bifurcations*, Trans Amer. Math. Soc., 333 (1992), pp. 397–422.

[Mi1] K. Mischaikow, *Homoclinic orbits in Hamiltonian systems and heteroclinic orbits in gradient and gradient-like systems*, J. Differential Equations, 81 (1989), pp. 167–213.

[Mi2] ———, *Travelling waves for a cooperative and a competative-cooperative system*, in Viscous Profiles and Numerical Methods fro Shock Waves, M. Shearer, ed., SIAM, Philadelphia, 1991.

[M-H] K. Mischaikow and V. Hutson, *Travelling waves for mutualist species*, SIAM J. Math. Anal., 25 (1993), pp. 987–1008.

[M-R] K. Mischaikow and J. Reineck, *A product theorem for connection matrices and the structure of connecting orbits*, JNA-TMA, to appear.

[Mo] R. Moeckel, *Morse decompositions and connection matrices*, Ergodic Theory Dynamical Systems, 8 (1988), pp. 227–249.

[Re1] J. Reineck, *The connection matrix in Morse–Smale flows*, Trans. Amer. Math. Soc., 322 (1990), pp. 523–544.

[Re2] ———, *A connection matrix analysis of ecological models*, JNA-TMA, 17 (1991), pp. 361–384.

[Sa] D. SALAMON, *Connected simple systems and the Conley index of isolated invariant sets*, Trans. Amer Math. Soc., 291 (1985), pp. 1–41.

[Sm] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

[Te] D. TERMAN, *Directed graphs and travelling waves*, Trans. Amer Math. Soc., 289 (1985), pp. 809–847.

SIAM J. MATH. ANAL.
Vol. 24, No. 5, pp. 1215–1225, September 1993

005

SINGULARITIES OF THE X-RAY TRANSFORM AND LIMITED DATA TOMOGRAPHY IN $\mathbb{R}^2$ AND $\mathbb{R}^3$*

ERIC TODD QUINTO†

Abstract. Given a function f, the author specifies the singularities of f that are visible in a stable way from limited X-ray tomographic data. This determines which singularities of f can be stably recovered from limited data and which cannot, no matter how good the inversion algorithm. Microlocal analysis is used to determine the relationship between the singularities of a function f and those of its X-ray transform. The results are applied to determine the singularities that are visible for limited angle tomography and the interior and exterior problems. The author also suggests a practical method to use this relationship to reconstruct singularities of f from limited data Rf. The X-ray transform with sources on a curve in $\mathbb{R}^3$ is also analyzed.

Key words. X-ray transform, limited data tomography, microlocal analysis

AMS subject classifications. primary 44A12, 92C55; secondary 35S30, 58G15

1. Introduction. X-ray tomography is an important, noninvasive, practical way of finding the density of objects. In standard tomography, X-rays of the object are taken over an evenly distributed set of lines, so-called *complete tomographic data*, and well-known algorithms are used to recover a good approximation to that object [21]. Inversion is only mildly ill-conditioned (continuous of order $\frac{1}{2}$ in Sobolev norms). However, one often needs to find the density of an object but one cannot get X-ray tomographic data over an evenly spaced set of lines through the object but only some subset; one has *limited tomographic data.* Limited data tomography is important in medical imaging [21], scientific tomography [1], and industrial nondestructive evaluation [28].

In general, reconstruction from limited tomographic data is much more highly ill-posed than reconstruction from complete data [6]. As a result, inversion algorithms using limited data, generally, can create artifacts, blurring or other distortions in their reconstructions. The goal of this article is to classify what singularities can be stably reconstructed from limited data and what singularities cannot be stably reconstructed no matter how good the algorithm. To do this, we will use a precise concept of singularity: *the wavefront set,* and a precise concept of stability: *continuity in microlocal Sobolev norms.* Then we will tell which singularities the X-ray transform "sees" stably and which singularities are not stably detected from limited data. The reason we can do this is because the X-ray transform is an elliptic Fourier integral operator and, therefore, changes wavefront sets in specific ways.

We do not claim that all limited data tomography algorithms will reconstruct the "visible" singularities well. Rather, we claim that, if a singularity is not stably visible from limited data, no algorithm can reconstruct it stably. For "visible" singularities our theorem gives stability estimates of order $\frac{1}{2}$ in Sobolev norms, so one would ex-

*Received by the editors July 13, 1992; accepted for publication (in revised form) February 8, 1993.

† Department of Mathematics, Tufts University, Medford, Massachusetts 02155 (equinto@math.tufts.edu). The author was partially supported by the Emmy Noether Institute at Bar Ilan University, the Humboldt Stiftung, and National Science Foundation grants MCS 8901203 and MCS 9123862.

pect "visible" singularities to be well constructed by a "good" algorithm even in the presence of noise.

This work is a natural outgrowth of [26], which gave the general principle (3.3) we make more precise and then prove in §3–4. Palamodov stated a closely related idea in [22]. The "tangent casting" effects of [30] is an intuitive way of expressing (3.3) below. One can also understand stability of these problems using singular value decompositions [4], [14], [16], [17], [18]. Lambda tomography [5] is a well developed algorithm that finds singularities of a function from real tomographic data. Their method works quite well with interior data (Example 3.3). Ramm and Zaslavsky [29] have developed a method using Legendre transforms to reconstruct the singularities of a function from knowing the singularities of its Radon transform. They consider functions $f = \psi\chi_D$, where D is a piecewise smooth domain and ψ is smooth—functions with the jump singularities on ∂D, and they use the behavior of Rf at ∂D to find the (jump) singularities of f. Technicians currently use the sinogram, the graph of $Rf(\theta, p)$ in rectangular coordinates, to find boundaries, but this method is subjective.

In Remark 3.2, we propose a method to reconstruct singularities (classified by Sobolev wavefront set) for arbitrary functions from general limited data.

Section 2 of this article provides the definitions of singularity and microlocal Sobolev smoothness. In §3 we give the singularity result for the Radon transform in the plane, Theorem 3.1. We apply this to determining singularities of arbitrary functions from general limited data (Remark 3.2) and to show limitations inherent in the common types of limited data tomography (Examples 3.3–3.5). Reconstructions from exterior data are presented that illustrate our analysis. §4 gives analysis and results for the X-ray transform with sources on a curve in $\mathbb{R}^3$.

2. Microlocal singularities and Sobolev spaces. Our development is valid for distributions as well as functions, so first, we recall some basic definitions. $\mathcal{D}(\mathbb{R}^n)$ is the space of C^∞ functions of compact support. A distribution $f \in \mathcal{D}'(\mathbb{R}^n)$ is a continuous linear functional on $\mathcal{D}(\mathbb{R}^n)$. A distribution f has compact support if there is a compact set $K \subset \mathbb{R}^n$ such that $f(\phi) = 0$ for all functions $\phi \in \mathcal{D}(\mathbb{R}^n)$ with support disjoint from K, that is, f is zero outside of K. The set of distributions of compact support is denoted by $\mathcal{E}'(\mathbb{R}^n)$.

The wavefront set of $f \in \mathcal{D}'(\mathbb{R}^n)$ is a powerful classification of singularities because it involves not only a point x_0 at which f is not smooth, but also a *direction* in which f is not smooth at x_0. To understand this we recall some facts about the Fourier transform. When f has compact support, then f is equal to a C^∞ function almost everywhere if and only if its Fourier transform, $\mathcal{F}f$, decreases rapidly in all directions (for all $N \in \mathbb{N}$, there exists C_N such that for all $\xi \in \mathbb{R}^n$, $|\mathcal{F}f(\xi)| \le C_N(1 + |\xi|)^{-N}$). This relates global smoothness of f to rapid decrease of its Fourier transform. A local version of this at a point $x_0 \in \mathbb{R}^n$ would be obtained by multiplying f by a smooth cut-off function, ϕ (with $\phi(x_0) \neq 0$) and seeing if this Fourier transform is rapidly decreasing in every direction. However, this localized Fourier transform $\mathcal{F}(\phi f)$ gives even more specific information—microlocal information—namely, the *directions* in which $\mathcal{F}(\phi f)$ does not decrease rapidly.

DEFINITION 2.1. Let $f \in \mathcal{D}'(\mathbb{R}^n)$ and let $x_0 \in \mathbb{R}^n$ and $\xi_0 \in \mathbb{R}^n \setminus 0$. Then we say $(x_0, \xi_0) \in \mathrm{WF}f$, the wavefront set of f, if and only if for each cut-off function at x_0, $\phi \in \mathcal{D}(\mathbb{R}^n)$ with $\phi(x_0) \neq 0$, $\mathcal{F}(\phi f)$ does not decrease rapidly in any open conic neighborhood of the ray $\{t\xi_0 | t > 0\}$.

For example, if $D \subset \mathbb{R}^n$ has smooth boundary, then $\mathrm{WF}\chi_D$ is exactly the set of normals to ∂D. One can prove this using the definition and a local coordinate change to flatten ∂D locally. If ψ is a smooth function then $\mathrm{WF}\psi\chi_D \subset \mathrm{WF}\chi_D$, and if ψ is not zero anywhere on ∂D, then these sets are equal [11], [24, Lemma 13.3, p. 279].

As defined, WFf is a closed subset of $\mathbb{R}^n \times (\mathbb{R}^n \setminus 0)$ that is conic in the second variable. The Sobolev space analogue to the concept of microlocal smoothness is as follows (see [24, p. 259]).

DEFINITION 2.2. A distribution f is in the Sobolev space H^s microlocally near (x_0, ξ_0) if and only if there is a cut-off function $\phi \in \mathcal{D}(\mathbb{R}^n)$ with $\phi(x_0) \neq 0$ and function $u(\xi)$ homogeneous of degree zero and smooth on $\mathbb{R}^n \setminus 0$ and with $u(\xi_0) \neq 0$ such that $u(\xi)\mathcal{F}(\phi f)(\xi) \in L^2(\mathbb{R}^n, (1 + |x|^2)^s)$.

First, one localizes near x_0 by multiplying f by ϕ and then one takes Fourier transform. Finally, one *microlocalizes* near ξ_0 by forming $u\mathcal{F}f$ and see if this is in $\mathcal{F}(H^s(\mathbb{R}^n))$. It follows from the definition that, if $(x_0, \xi_0) \notin \mathrm{WF}f$, then for all s, f is H^s near (x_0, ξ_0).

Wavefront set and microlocal smoothness are usually defined on $T^*(\mathbb{R}^n) \setminus 0$, the cotangent space of $\mathbb{R}^n$ with its zero section removed, because such a definition can be extended invariantly to manifolds using local coordinates. For the manifold $[0, 2\pi] \times \mathbb{R}$, choosing a function $\phi(\theta, p)$ with sufficiently small support allows one to use θ and p as local coordinates. We will use these conventions.

3. The X-ray transform in the plane.

In $\mathbb{R}^2$ the microlocal analysis of the X-ray transform is easier to describe if one uses parallel-beam geometry rather than fan-beam geometry. By rebinning—a coordinate change—the results are the same as for fan-beam data for functions supported inside the circle of sources. First, let $\cdot$ denote the standard inner product on $\mathbb{R}^2$; let $|\ |$ be the induced norm. Let $\theta \in [0, 2\pi]$, and let $p \in \mathbb{R}$. Let $\overline{\theta} = (\cos\theta, \sin\theta)$ and $\theta^\perp = (-\sin\theta, \cos\theta)$. Now let $\ell(\theta, p) = \{x \in \mathbb{R}^2 | x \cdot \overline{\theta} = p\}$, the line with normal vector $\overline{\theta}$ and directed distance p from the origin. The points (θ, p) and $(\theta + \pi, -p)$ parameterize the same line $\ell(\theta, p)$. Let ds be arc length, the measure on $\ell(\theta, p)$ induced from Lebesgue measure on $\mathbb{R}^2$. The classical X-ray transform in the plane is defined for an integrable function f on $\mathbb{R}^2$ by

$$Rf(\theta, p) = \int_{y \in \ell(\theta, p)} f(y)\,ds. \tag{3.1}$$

$Rf(\theta, p)$ is the integral of f over the line $\ell(\theta, p)$.

In order to describe the main theorem, we will consider wavefront sets as subsets of cotangent spaces. To this end, let $x_0 \in \mathbb{R}^2$. If $y = (y_1, y_2) \in \mathbb{R}^2$, then we let $y\mathbf{dx} = y_1\mathbf{dx}_1 + y_2\mathbf{dx}_2$ be the cotangent vector corresponding to y in $T^*_{x_0}\mathbb{R}^2$. Now let $(\theta_0, p_0) \in [0, 2\pi] \times \mathbb{R}$. Here we will identify $[0, 2\pi]$ with the unit circle by equating zero with 2π. Then for $(\theta, p) \in [0, 2\pi] \times \mathbb{R}$, we let $\mathbf{d\theta}$ and $\mathbf{dp}$ be the standard basis of $T^*_{(\theta,p)}([0, 2\pi] \times \mathbb{R})$. The theorem follows.

THEOREM 3.1. *Let $f \in \mathcal{E}'(\mathbb{R}^2)$. If $(x; \xi) \in T^*(\mathbb{R}^n) \setminus O$ is not conormal to $\ell(\theta_0, p_0)$, then wavefront set of f at $(x; \xi)$ does not contribute to* WFRf *above (θ_0, p_0). Let $x_0 \in \ell(\theta_0, p_0)$ and let $\eta_0 = \mathbf{dp} - (x_0 \cdot \theta_0^\perp)\mathbf{d\theta}$. Let $a \neq 0$. The correspondence between* WFf *and* WFRf *is:*

$$(x_0; a\overline{\theta}_0\mathbf{dx}) \in \mathrm{WF}f \quad \textit{if and only if } (\theta_0, p_0; a\eta_0) \in \mathrm{WF}Rf. \tag{3.2}$$

Given $(\theta_0, p_0; a\eta_0)$, $(x_0; a\overline{\theta}_0\mathbf{dx})$ is uniquely determined by (3.2). *Moreover, $f \in H^s$ is microlocally near $(x_0; a\overline{\theta}_0\mathbf{dx})$ if and only if $Rf \in H^{s+1/2}$ is microlocally near $(\theta_0, p_0; a\eta_0)$.*

Theorem 3.1 provides an exact correspondence between singularities of f and those of Rf. Moreover, it states that the singularities of Rf that are detected are of Sobolev order $\frac{1}{2}$ smoother than the corresponding singularities of f. For typical singularities of f (jump singularities in $H^{1/2-\epsilon}$) one can realistically expect the corresponding singularities of Rf not to be masked by noise. Reconstructions given in Figs. 1 and 2 will corroborate this.

The theorem has this simple corollary:

(3.3)
The X-ray transform data Rf for (θ, p) arbitrarily near (θ_0, p_0) detects singularities of f perpendicular to the line $\ell(\theta_0, p_0)$ but not in other directions.

This follows because of the correspondence (3.2): $Rf(\theta, p)$ is smooth near (θ_0, p_0) (no wavefront set near this point) if and only if there is no wavefront set of f at points on $\ell(\theta_0, p_0)$ conormal to the line.

As an example, let D be a compact set with smooth boundary and let $f = \psi\chi_D$, where ψ is a smooth function that is not zero anywhere on ∂D. Then, by (3.3), $Rf(\theta, p)$ is smooth near (θ_0, p_0) if and only if $\ell(\theta_0, p_0)$ is not tangent to ∂D. If ∂D is not smooth then more wavefront directions will appear at points where ∂D is not smooth. Remark 3.2 gives a more general observation with practical implications.

Remark 3.2. The correspondence (3.2) gives a way to find WFf from knowing WFRf. Given $(\theta_0, p_0; a\eta_0) \in \mathrm{WF}Rf$, the rule (3.2) determines $(x_0; a\overline{\theta}_0\mathbf{dx})$ uniquely. This method to find singularities of f is easiest to describe in the case f is C^1 except for jump singularities on a collection, E, of C^1 curves. In this case, almost all singularities of f are in $H^{1/2-\epsilon}$ (so corresponding singularities of Rf are in $H^{1-\epsilon}$) for $\epsilon > 0$ but not for $\epsilon = 0$. One can take a local (discrete) Fourier transform of Rf in (θ, p) and find the directions in which the localized transform is not in $\mathcal{F}H^1$. Perhaps this can be efficiently done just by calculating local fast Fourier transforms and looking for directions in which they do not decrease quickly. Then the rule (3.2) gives the covectors $(x_0; a\overline{\theta}_0\mathbf{dx})$ at which f is not $H^{1/2}$. These covectors specify the jump singularities of f, that is the location of E (and $\overline{\theta}_0$ even gives the normal to E at x_0). This method also filters out noise that is H^1 or smoother.

This method can be used for limited data problems: the method is local in the strong sense that singularities of Rf at (θ_0, p_0) (and therefore the corresponding singularities of f on $\ell(\theta_0, p_0)$) are determined by data $Rf(\theta, p)$ for (θ, p) near (θ_0, p_0). This method is being pursued.

Proof of Theorem 3.1. The microlocal correspondence between WFf and WFRf is in the literature (e.g., [8], [25]), but since it is especially straightforward in this case, it will be given here. First, note that the Schwartz kernel of the operator R is the distribution on $\mathbb{R}^2 \times \big([0, 2\pi] \times \mathbb{R}\big)$ that is integration with respect to the weight $dxd\theta$ over the set $Z = \{(x, \theta, p) | x \cdot \overline{\theta} = p\}$. This is a special type of distribution and in [10] it is shown to be a Fourier integral distribution associated with the Lagrangian manifold $\Gamma = N^*Z \setminus 0$ where N^*Z is the conormal bundle of Z in $T^*\big(\mathbb{R}^2 \times ([0, 2\pi] \times \mathbb{R})\big)$. As shown in [8] (see also [25] for details), because the measure of integration $dxd\theta$ is nowhere zero and the projection from Γ to $T^*([0, 2\pi] \times \mathbb{R}) \setminus 0$ is a injective immersion, R is elliptic with elliptic inverse that composes well with R. To understand what R does to wavefront sets, one must calculate the set Γ. Z is defined by the equation $x \cdot \overline{\theta} - p = 0$ and so its differential, $\overline{\theta}\mathbf{dx} + x \cdot \theta^\perp \mathbf{d\theta} - \mathbf{dp}$, is a basis of N^*Z at each point. Therefore,

$$\Gamma = \big\{\big(x, \theta, p; a(\overline{\theta}\mathbf{dx} + x \cdot \theta^\perp \mathbf{d\theta} - \mathbf{dp})\big) \big| (x, \theta, p) \in Z,\ a \neq 0\big\}. \tag{3.4}$$

By the calculus of elliptic Fourier integral operators, there is a simple correspondence between WFf and WFRf: $(x; \xi) \in \mathrm{WF}f$ if and only if there is a $(\theta, p; \eta) \in \mathrm{WF}Rf$

with $(x, \theta, p; \xi, -\eta) \in \Gamma$ [31]. Using (3.4) we see this correspondence is exactly (3.2). Furthermore this correspondence coming from (3.4) shows that if $(x; \xi)$ is not conormal to $\ell(\theta_0, p_0)$, then wavefront set of f at $(x; \xi)$ does not contribute to WFRf above (θ_0, p_0). To see that (3.2) uniquely determines $(x_0; \overline{\theta}_0 \mathbf{dx})$, first note that a is determined by the $\mathbf{dp}$ coordinate of $a\eta_0$. Then as $a \neq 0$, $x_0 \cdot \theta_0^\perp$ is determined by the $\mathbf{d\theta}$ coordinate of $a\eta_0$, and finally $x_0 \cdot \overline{\theta}_0 = p$ determines x_0.

The assertion about H^s will be given because, although it is straightforward, it is not in the elementary literature. We prove one direction and leave the other to the reader. Let Rf be in $H^{s+1/2}$ near $(\theta_0, p_0; a\eta_0)$. Then, by Theorem 6.1 [24, p. 259], $Rf = u_1 + u_2$ where $u_1 \in H_c^{s+1/2} = H^{s+1/2} \cap \mathcal{E}'$, $u_2 \in \mathcal{E}'$, and $(\theta_0, p_0; a\eta_0) \notin \text{WF}u_2$. Because R^{-1} is a Fourier integral operator continuous of order $\frac{1}{2}$ [8] and $u_1 \in H_c^{s+1/2}$, $R^{-1}u_1 \in H_{\text{loc}}^s(\mathbb{R}^2)$ [31]. R^{-1} is a Fourier integral operator associated to Γ (with $\mathbb{R}^2$ and $[0, 2\pi] \times \mathbb{R}$ coordinates reversed) and so the "inverse" relation to (3.2) holds for R^{-1}. Therefore, as $(\theta_0, p_0; a\eta_0) \notin \text{WF}u_2$, $(x_0; a\overline{\theta}_0 \mathbf{dx}) \notin \text{WF}(R^{-1}u_2)$. Therefore, $f = R^{-1}u_1 + R^{-1}u_2$ is the sum of a distribution in H_{loc}^s and one that is smooth near $(x_0; a\overline{\theta}_0 \mathbf{dx})$. Therefore, by Theorem 6.1 [24, p. 259], f is microlocally H^s near $(x_0; a\overline{\theta}_0 \mathbf{dx})$. □

We now apply Theorem 3.1 and (3.3) to three common types of limited data tomography in the plane.

Example 3.3. *Limited angle tomography.* Let $U \subset [0, 2\pi]$ be open, $U = U + \pi$ mod 2π. In limited angle tomography, one knows data $Rf(\theta, p)$ for all p and for $\theta \in U$. One can reconstruct $f(x)$ for all $x \in \mathbb{R}^2$ from limited angle data [21]. However, by (3.3), the only singularities of f that one can detect in a stable way are those with directions in U. To see this, choose $x \in \mathbb{R}^2$ and $\theta \in U$. Any wavefront of f at $(x; \overline{\theta}\mathbf{dx})$ is detected by limited angle data because the line $\ell(\theta, x \cdot \overline{\theta})$ is in this data set. For the same reason, wavefront of f at $(x; \overline{\theta}\mathbf{dx})$ for $\theta \notin U$ will not be stably detected by this limited angle data.

This phenomenon is illustrated by the singular functions in [14]. Those corresponding to large singular values (easy to reconstruct) oscillate generally in directions in U and those corresponding to small singular values (hard to reconstruct) oscillate generally in directions outside of U. This is also seen in the actual reconstructions from limited angle tomography.

Example 3.4. *The interior problem.* Let $M > 1$ and assume supp $f \subset \{x \in \mathbb{R}^2 \big| |x| \leq M\}$. In this problem, one has data $Rf(\theta, p)$ for all θ but only for $|p| < 1$. The goal is to reconstruct $f(x)$ for $|x| < 1$. Simple examples show this is impossible in general. However, according to (3.3), one can detect all singularities of f in $|x| < 1$. To see this, choose a point x inside the unit disk and choose a direction $\theta \in [0, 2\pi]$. Then the line through x and normal to $\overline{\theta}$ is in the data set for interior tomography and so any singularity of f at $(x; \overline{\theta}\mathbf{dx})$ is detected by interior data.

Lambda tomographic reconstructions are local—they use data $Rf(\theta, p)$ only for lines $\ell(\theta, p)$ near x to reconstruct at x. So Lambda tomography is useful for the interior problem. In fact, Lambda tomographic reconstructions for the interior problem clearly show the singularities of f in the unit disk [5]. Maass [18] has developed a singular value decomposition for this problem. See also [16].

Example 3.5. *The exterior problem.* Assume supp $f \subset \{x \in \mathbb{R}^2 \big| |x| \leq M\}$. Here one has data $Rf(\theta, p)$ for all θ but only for $|p| > 1$. By [3] one can reconstruct $f(x)$ for $|x| > 1$. Let $|x| > 1$ and $\theta \in [0, 2\pi]$. Then the only singularities of f at x that are reconstructed in a stable manner are those for θ with $\ell(\theta, x \cdot \overline{\theta})$ in the data set, that is, for $|x \cdot \overline{\theta}| > 1$. Other singularities of f are not stably detected. This can be seen from the reconstructions in Figs. 1 and 2.

Lewitt and Bates [13], Louis [15], and Natterer [20] have developed good reconstruction algorithms that use exterior data. The author has developed an exterior reconstruction algorithm which employs Perry's singular value decomposition [23] and a priori information about the shape of the object to be reconstructed. Reconstructions for "medical" phantoms are in [26] and those for industrial phantoms are in [27], [28]. Exactly those singularities that are supposed to be stably reconstructed are clearly defined. In the author's algorithm, singularities that are not "visible" are smeared; reconstructions will now be given.

Figure 1 shows an object with outer radius $M = 1.5$ on a rectangular grid. The two bigger circles have density 1.5 and the two smaller 1.375. The annulus has density one. The reconstruction in Fig. 1b is gotten using the author's algorithm with noiseless data. The reconstruction in Fig. 1c is from the same algorithm but using data with slightly less than 1% L^∞ noise. Data are taken over lines using 100 values of p and 256 of θ [26]. In both reconstructions, the principle (3.3) is illustrated. However, the reconstruction with noise, Fig. 1c, shows some algorithm limitations as well (and reminds one that algorithm limitations independent of the principle can be important). The slightly darker areas in the background in Fig. 1c are the result of amplified high polar Fourier coefficients due to the noisy data.

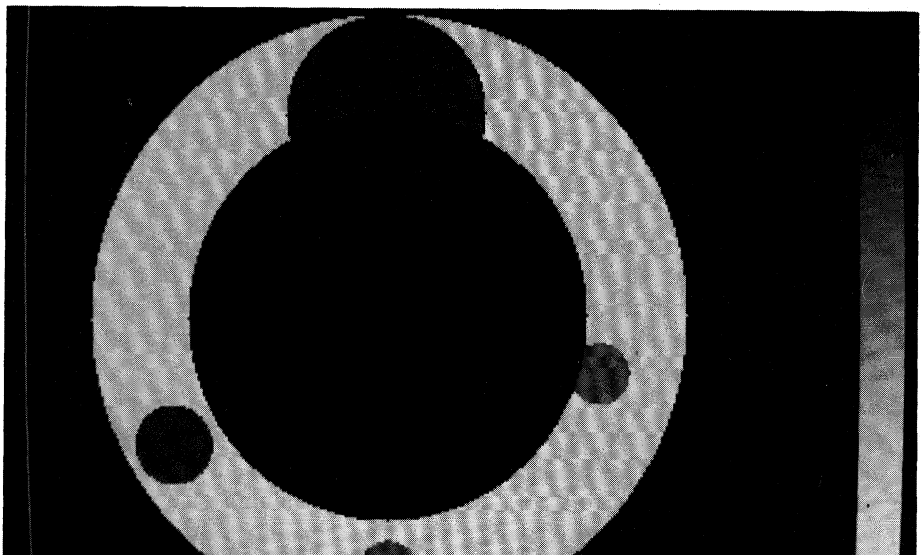


FIG. 1a. *Rectangular coordinate display of the phantom* [26] (*similar to the phantom in* [20]).

Figure 2 shows polar coordinate displays for $x = (r, \theta)$ with $\theta \in [\pi/8, 3\pi/8]$ on the horizontal axis and $r \in [1, 1.10]$ on the vertical axis ($r = 1.10$ at the bottom) [28]. To provide sufficiently fine radial resolution, the scale in r is magnified by a factor of 7.85. The phantom in Fig. 2 is supposed to represent a rocket motor with fuel of density 1.7 inside the circle of radius $r = 1.052$, an insulator of density 1.1 from

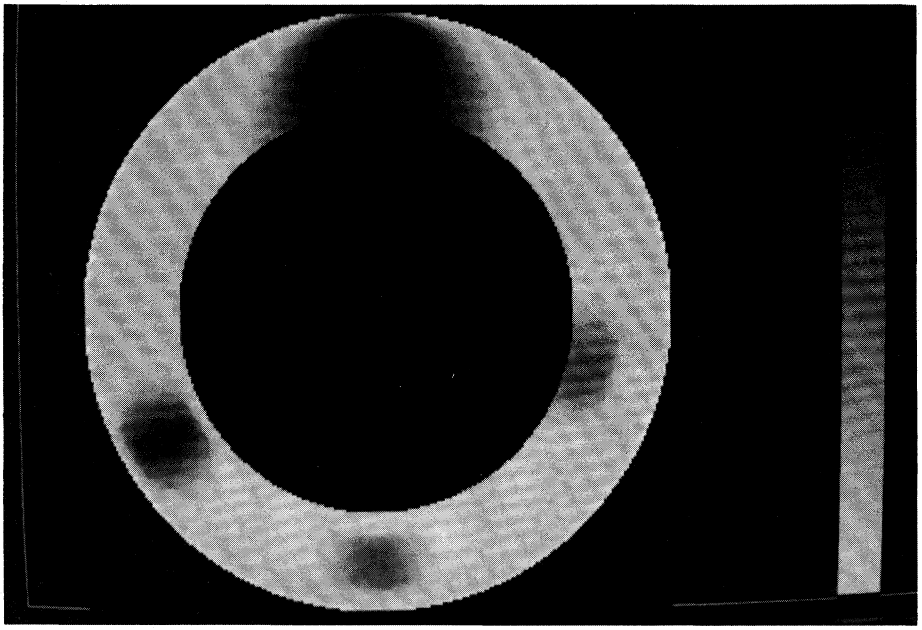


FIG. 1b. *Rectangular coordinate display of the reconstruction without noise of the phantom in Fig.* 1a [26].

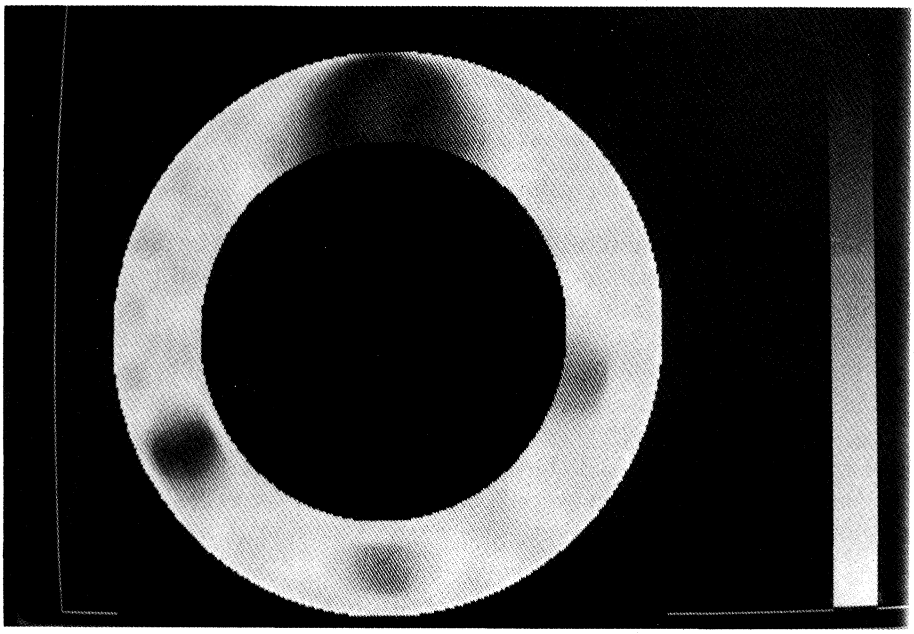


FIG. 1c. *Rectangular coordinate display of the reconstruction with noise of the phantom in Fig.* 1a [26].

$1.052 < r < 1.056$, and a shell of density 1.5 and outer radius $r = 1.093$. The defect rests against the inside boundary of the insulator and extends for 0.06 radians and is 0.0014 units thick (it is seen tangentially by only three detectors). It is centered at $\pi/4$ radians and has density zero. The reconstruction is done with 1% multiplicative L^∞ noise. Data are collected in a fan beam with 200 rays from $p = 1.0$ to $p = 1.10$ that emanate from the source in evenly spaced angles. The source and fan beam rotate around the object in 512 equally spaced angles.

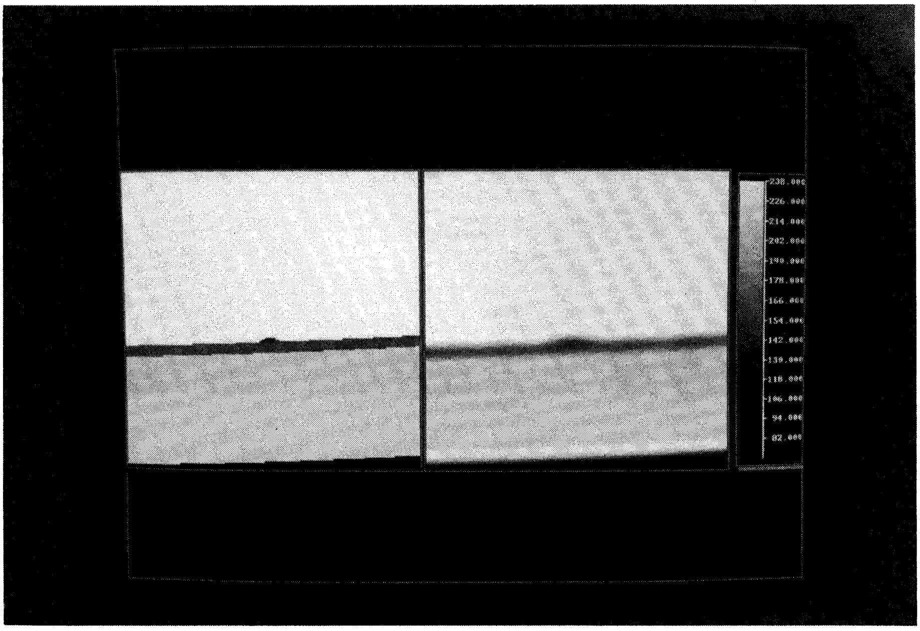


FIG. 2. *Polar coordinate display of rocket motor (phantom left, reconstruction with noise right)* [28]. *The "wedge" near* $r = 1.10$ *(at the bottom of the display) occurs because the center of the rocket is offset slightly from the center of the coordinate system.*

Because some wavefront directions are not stably detectable by limited angle data or by exterior data, inversion for these problems is highly ill-posed (see the example in [6] and the inverse discontinuity result in L^2 of [19]).

Reconstructions in Figs. 1 and 2 illustrate the principle (3.3). In the reconstruction in Fig. 1, the "sides" of the circles are blurred (corresponding to singularities normal to lines *not* in the data set), but the "inside" and "outside" boundaries of the circles are well reconstructed. This occurs despite the fact that only a few lines in the data set are tangent to the inside boundaries. The reconstructions of Fig. 2 are good (even though the problem is, in general, highly ill-posed) because all singularities are perpendicular to lines in the data set. This is true, even though the defect is very thin and short in extent.

4. The X-ray transform with sources on a curve in $\mathbb{R}^3$. The standard parameterization will be used for the divergent beam transform in $\mathbb{R}^3$. Let $\omega \in S^2$ and $x \in \mathbb{R}^3$, then the ray $\mathfrak{r}(\omega, x) = \{x + t\omega | t \geq 0\}$ is the ray parallel to ω and starting at x. The divergent beam transform of $f \in C_c(\mathbb{R}^3)$ is

$$Df(\omega, x) = \int_{t=0}^{\infty} f(x + t\omega)dt, \tag{4.1}$$

the integral of f over the ray $\mathfrak{r}(\omega, x)$. Typically, the sources for the divergent beam transform are points on a smooth closed curve γ. The divergent beam transform is defined for $f \in L^1_c(\mathbb{R}^3 \setminus \gamma)$ (L^1 functions of compact support in $\mathbb{R}^3 \setminus \gamma$) [9] (and even continuous on $\mathcal{E}'(\mathbb{R}^3 \setminus \gamma)$, [7]).

Inversion of the divergent beam transform is a limited data problem because data are given only over rays with sources on γ. Moreover, typically, X-rays are taken only over an open connected set, $\mathfrak{R}$, of rays with sources on γ. In general, as long as some ray in the data set is disjoint from supp f, then the part of f seen by the data (that is, supp $f \cap \left[\cup_{\mathfrak{r}\in\mathfrak{R}} \mathfrak{r}\right]$) is uniquely determined (see [9] and the generalization [2, Thm. 2.2]). Our theorem for the X-ray transform is as follows.

THEOREM 4.1. *Let γ be a smooth curve in $\mathbb{R}^3$ and $f \in \mathcal{E}'(\mathbb{R}^3 \setminus \gamma)$. Let $x_0 \in$ supp f and $\xi_0 \in T^*_{x_0}(\mathbb{R}^3) \setminus 0$. Then any wavefront set of f at $(x_0; \xi_0)$ is stably detected from data Df with sources on γ if and only if*

(4.2) *the plane $\mathcal{P}$, through x_0 conormal to ξ_0, intersects γ transversally.*

If data are taken over an open set of rays with sources on γ, then a ray in $\mathcal{P}$ from γ to x_0 must be in the data set for (4.2) *to apply. In these cases, f is in H^s microlocally near $(x_0; \xi_0)$ if and only if the corresponding singularity of Df is in $H^{s+1/2}$.*

The exact correspondence of singularities analogous to (3.2) can be obtained from the microlocal diagram (3.1.1) and the proof of Proposition 3.1.1 of [2]. Theorem 4.1 follows from [7] as well.

The global version of condition (4.2)—every plane meeting supp f intersects γ transversally—is called the Kirillov–Tuy condition. This condition is required for the inversion methods of Kirillov [12] and Tuy [32]. Under this condition, Finch [6] proves that $f \in H^s$ if $Df \in H^{s+1/2}$ for $s \geq \frac{1}{2}$ (and our theorem implies this fact for all s).

Typically, X-ray sources are placed on a circle surrounding the object to be reconstructed. Theorem 4.1 shows the singularities that are not detected by such data: singularities $(x_0; \xi_0)$ conormal to planes $\mathcal{P}$ not meeting the circle transversally. There are many such singularities—the more undetected singularities, the farther x_0 is from the plane of C. Finch [6] and others have noted that inversion with sources on one circle, C, is highly unstable. By analyzing the singular values, Maass shows that inversion is more stable for nonplanar curves such as two parallel circles or curves oscillating on a cylinder [17]. Condition (4.2) is another way to understand why inversion of data with sources on such nonplanar curves is better posed than for sources on one circle—in general, if the curve is nonplanar, more singularities can be detected stably from the given data.

Proof of Theorem 4.1. The microlocal assertion of Theorem 4.1 is a paraphrase of the comment about "type II complexes" below the statement of Proposition 3.1.1 of [2]. That comment is equivalent to the fact that if $x_0 \in \mathfrak{r}(\omega, y)$ for some $y \in \gamma$ and ξ_0 is conormal to ω then WFf at $(x_0; \xi_0)$ is detected by divergent beam data unless ξ_0 is conormal to γ at y. This is equivalent to condition (4.2). The statement about microlocal Sobolev spaces is valid because D is an elliptic Fourier integral operator of order $-\frac{1}{2}$ and so, if a singularity of f is detected by data Df, then the singularity of Df is $\frac{1}{2}$ order smoother than thę corresponding singularity of f. This can be proven just as the analogous statement in Theorem 3.1 is proven. □

Acknowledgments. The author thanks Bar Ilan University and Universität des Saarlandes for their hospitality as this article was being written. He is indebted to A. Zaslavsky for lively and informative conversations about his article with A. G.

Ramm [29]. He is also indebted to the many conversations with Adel Faridani, Alfred Louis, Peter Maass, Frank Natterer, Kennan Smith, and others over the years that have helped develop and refine the ideas in this article. Finally, the author thanks the Institut für Numerische und instrumentelle Mathematik der Universität Münster for the display programs for Fig. 2.

REFERENCES

[1] M. D. Altschuler, *Reconstruction of the global-scale three-dimensional solar corona*, in Image Reconstruction From Projections, Implementation And Applications, G. T. Herman, ed., Topics Appl. Phys., 32, Springer-Verlag, New York, Berlin, pp. 105–145.

[2] J. Boman and E. T. Quinto, *Support theorems for real-analytic Radon transforms on line complexes in three-space*, Trans. Amer. Math. Soc., 335 (1993), pp. 877–890.

[3] A. M. Cormack, *Representation of a function by its line integrals with some radiological applications*, J. Appl. Phys., 34 (1963), pp. 2722–2727.

[4] M. Davison, *The ill-conditioned nature of the limited angle tomography problem*, SIAM J. Appl. Math., 43 (1983), pp. 428–448.

[5] A. Faridani, E. L. Ritman, and K. T. Smith, *Local tomography*, SIAM J. Appl. Math., 52 (1992), pp. 459–484.

[6] D. V. Finch, *Cone beam reconstruction with sources on a curve*, SIAM J. Appl. Math., 45 (1985), pp. 665–673.

[7] A. Greenleaf and G. Uhlmann, *Non-local inversion formulas in integral geometry*, Duke J. Math., 58 (1989), pp. 205–240.

[8] V. Guillemin and S. Sternberg, *Geometric Asymptotics*, Amer. Math. Soc., Providence, RI, 1977.

[9] C. Hamaker, K. T. Smith, D. C. Solmon, and S. L. Wagner, *The divergent beam X-ray transform*, Rocky Mountain J. Math., 10 (1980), pp. 253–283.

[10] L. Hörmander, *Fourier integral operators* I, Acta Math. 127 (1971), pp. 79–183.

[11] ———, *The Analysis of Linear Partial Differential Operators* I, Springer-Verlag, New York, 1983.

[12] A. A. Kirillov, *On a problem of I. M. Gel'fand*, Soviet Math. Dokl., 2 (1961), pp. 268–269.

[13] R. M. Lewitt and R. H. T. Bates, *Image reconstruction from projections.* II: *Projection completion methods* (*theory*), Optik, 50 (1978), pp. 189–204; III: *Projection completion methods* (*computational examples*), Optik, 50 (1978), pp. 269–278.

[14] A. Louis, *Incomplete data problems in X-ray computerized tomography* 1. *Singular value decomposition of the limited angle transform*, Numer. Math., 48 (1986), pp. 251–262.

[15] ———, private communication.

[16] A. Louis and A. Rieder, *Incomplete data problems in X-ray computerized tomography* II. *Truncated projections and region-of-interest tomography*, Numer. Math., 56 (1989), pp. 371–383.

[17] P. Maass, 3D *Röntgentomographie: Ein Auswahlkriterium für Abtastkurven*, Z. Angew. Math. Mech., 68 (1988), pp. T498–T499.

[18] ———, *The interior Radon transform*, SIAM J. Appl. Math., 52 (1992), pp. 710–724.

[19] W. Madych and S. Nelson, *Reconstruction from restricted Radon transform data: resolution and ill-conditionedness*, SIAM J. Math. Anal., 17 (1986), pp. 1447–1453.

[20] F. Natterer, *Efficient implementation of "optimal" algorithms in computerized tomography*, Math. Meth. Appl. Sci., 2 (1980), pp. 545–555.

[21] ———, *The mathematics of computerized tomography*, John Wiley, New York, 1986.

[22] V. Palamodov, *Nonlinear artifacts in tomography*, Sov. Phys. Dokl. 31 (1986), pp. 888–890.

[23] R. M. Perry, *On reconstruction of a function on the exterior of a disc from its Radon transform*, J. Math. Anal. Appl., 59 (1977), pp. 324–341.

[24] B. Petersen, *Introduction to the Fourier Transform and Pseudo-Differential Operators*, Pittman, Boston, 1983.

[25] E. T. Quinto, *The dependence of the generalized Radon transform on defining measures*, Trans. Amer. Math. Soc. 257 (1980), pp. 331–346.

[26] E. T. QUINTO, *Tomographic reconstructions from incomplete data–numerical inversion of the exterior Radon transform.*, Inverse Problems, 4 (1988), pp. 867–876.

[27] ———, *Limited data tomography in non-destructive evaluation*, Signal Processing Part II: Control Theory And Applications, IMA Vol. Math. Appl., 23, Springer-Verlag, New York, 1990, pp. 347–354.

[28] ———, *Computed tomography and Rockets*, in Mathematical Methods In Tomography, Proceedings, Oberwolfach, 1990, Lecture Notes in Math., 1497, Springer-Verlag, Berlin, New York, 1991, pp. 261–268.

[29] A. G. RAMM AND A. I. ZASLAVSKY, *Reconstructing singularities of a function from its Radon transform*, preprint.

[30] L. A. SHEPP AND S. SRIVASTAVA, *Computed tomography of PKM and AKM exit cones*, AT & T. Tech. J., 65 (1986), pp. 78–88.

[31] F. TREVES, *Introduction to Pseudodifferential and Fourier integral operators* II, Plenum Press, New York, 1980.

[32] H. K. TUY, *An inversion formula for cone-beam reconstruction*, SIAM J. Appl. Math., 43 (1983), pp. 546–552.

SIAM J. MATH. ANAL.
Vol. 24, No. 5, pp. 1226–1253, September 1993

006

HOMOGENIZATION OF DEGENERATE WAVE EQUATIONS WITH PERIODIC COEFFICIENTS*

YOUCEF AMIRAT[†], KAMEL HAMDACHE[‡], AND ABDELHAMID ZIANI[§]

Abstract. In this paper the authors discuss homogenization of hyperbolic equations involving periodic coefficients which are degenerate relative to a certain direction. The general scheme by which effective equations are obtained is as the reiterated homogenization. The first step of the process leads to equations describing the oscillatory behavior in the direction of the propagation. Next, space averaging in the degenerate direction gives the result. The process in the second step produces nonlocal effects.

Key words. homogenization, hyperbolic, oscillations, periodic media

AMS subject classifications. 35B25, 35B40, 35C10, 35L15

1. Introduction. The main purpose of this report concerns the study of the propagation and the interaction of oscillations for a class of degenerate wave equations in periodic microstructure. Many problems of wave propagation in composite media are modeled by equations of type: $\rho\,\partial_t^2 u - \operatorname{div}(k\,\mathrm{grad}\ u) + \theta\,\partial_t u = f$, where ρ is the fluid density, k is the bulk modulus of the medium, θ is the viscous damping coefficient, f refers to an external force, and u stands for the pressure. The coefficients ρ, k, and θ are highly oscillating, depending on two spatial variables usually referred to as macroscopic and microscopic variables. Let $\varepsilon > 0$ be the length scale which characterizes the heterogeneities of the medium. Under usual assumptions on ρ, k, and θ, discussions about effective equations can be found in Bensoussan, Lions, and Papanicolaou [8] and Sanchez-Palencia [18]. Our interest here is in the case in which the components of the anisotropic matrix k are taken to be negligible, i.e., small compared with ε^2, in certain direction of the domain, so that waves propagate only in the orthogonal direction. The matrix k is, therefore, of degenerate type, the classical homogenization results fail; see [7].

The outline of this paper is as follows. In §2, we discuss the equation

$$\rho\left(\frac{x}{\varepsilon},y,\frac{y}{\varepsilon}\right)\partial_t^2 u^\varepsilon - \operatorname{div}_x\left(k\left(\frac{x}{\varepsilon},y,\frac{y}{\varepsilon}\right)\mathrm{grad}_x u^\varepsilon\right) = f\left(t,x,\frac{x}{\varepsilon},y,\frac{y}{\varepsilon}\right) \tag{1.1}$$

for $t\in]0,T[$, $(x,y)\in\Omega$, where $T>0$, $\Omega=\mathbb{R}^n\times\mathcal{O}$, $\mathcal{O}$ is an open subset of $\mathbb{R}^m$, supplemented with initial conditions

$$u^\varepsilon(0,x,y)=\alpha(x,y),\quad \partial_t u^\varepsilon(0,x,y)=\beta(x,y),\quad (x,y)\in\Omega. \tag{1.2}$$

Here and below points in $\mathbb{R}^n$ are denoted by x and points in $\mathcal{O}$ are denoted by y. We denote by Y_k the unit cube in $\mathbb{R}^k$ and by $L^\infty_{\mathrm{per}}(Y_k)$ (respectively, $L^2_{\mathrm{per}}(Y_k)$) the space

*Received by the editors December 30, 1991; accepted for publication (in revised form) February 9, 1993.

[†] Laboratoire d'Analyse Numérique, Université Claude Bernard-Lyon 1, 43 boulevard du 11 Novembre 1918, 69622 Villeurbanne cedex, France.

[‡] CNRS-ENSTA/GHN, Centre de l'Yvette, chemin de la Hunière, 91120 Palaiseau, France.

[§] Département de Mathématiques, Université de Nantes, 2 rue de la Houssinière, 44072 Nantes cedex 03, France.

of functions $\phi \in L^\infty(\mathbb{R}^k)$ (respectively, $L^2_{\text{loc}}(\mathbb{R}^k)$) such that ϕ is Y_k-periodic. The local variables related to Y_n-periodicity (respectively, Y_m-periodicity) are denoted by ζ (respectively, η). Assuming some hypotheses regarding the data $\rho = \rho(x, \zeta, y, \eta)$, $k = k(\zeta, y, \eta)$ and $f = f(x, \zeta, y, \eta)$, we want to describe the asymptotic behavior of the solution u^ε of (1.1), (1.2), more precisely to derive the effective equation satisfied by the weak limit u of (u^ε) as ε goes to zero.

The homogenization process we use will be carried out through two distinct steps following the reiterated homogenization principle; see Bensoussan, Lions, and Papanicolaou [8]. The first one gives the behavior when considering oscillations $\zeta \in Y_n$ in the direction x of the propagation. The main tool is the method of oscillatory test functions introduced by Tartar [19], [20]. To a subsequence of solutions (u^ε) of (1.1), (1.2), one can associate a function $U = U(t, x, y, \zeta, \eta)$ which represents its oscillatory behavior. We shall use a characterization in periodic microstructure due to Nguetseng [16]. One can find other developments related to this subject in Allaire [2], E [10], and Nguetseng [17]. The representation $U = U(t, x, y, \eta)$, which does not depend on the microscopic variable ζ thanks to the compactness property in the direction x, is a solution of the family of wave equations, parametrized by (y, η):

$$(1.3) \qquad \tilde{\rho}(y,\eta)\, \partial_t^2 U - \operatorname{div}_x(\tilde{k}(y,\eta)\operatorname{grad}_x U) = \tilde{f}(t,x,y,\eta), \quad t \in]0,T[, \quad x \in \mathbb{R}^n,$$

$$(1.4) \qquad U(0,x,y,\eta) = \alpha(x,y), \quad \partial_t U(0,x,y,\eta) = \beta(x,y), \quad x \in \mathbb{R}^n.$$

Here $\tilde{\rho}(y,\eta) = \int_{Y_n} \rho(\zeta, y, \eta)\, d\zeta$, $\tilde{f}(t,x,y,\eta) = \int_{Y_n} f(t,x,\zeta,y,\eta)\, d\zeta$, and $\tilde{k}(y,\eta)$ is the homogenized matrix with respect to oscillations ζ for fixed y and η.

Since $u(t,x,y) = \int_{Y_m} U(t,x,y,\eta)\, d\eta$, the next step consists of averaging (1.3) with respect to η. By Laplace and Fourier transforms of (1.3), (1.4), we reduce the averaging problem to characterize some weak limits as in the study of parametrized families of hyperbolic problems discussed in [5]. We distinguish two cases. The first one describes the situation where $\tilde{k}$ is an isotropic tensor; we follow the pattern given in [3]. The result is obtained using an integral representation for Nevanlinna–Pick's holomorphic functions; see Ahiezer and Krein [1], Donoghue [9], and Amirat et al. [3]. For an anisotropic tensor $\tilde{k}$, we apply Radon transform to (1.3), (1.4) as discussed in [5]. This reduces the problem for $\mathbb{R}^n$ in a similar problem for $\mathbb{R}$. The resulting one-dimensional equation may be averaged within the framework mentioned above. Integration over the frequency domain together with use of inverse Radon transform give the homogenized problem. We mention here the strong interaction of the oscillations of the coefficients ρ and k as well as the source term f. The lack of compactness produces nonlocal effects described by integro-differential effective equations. Note that the choice of the dependence of the coefficients ρ and k in (1.1) is essential in the second step. To conclude this section, we address the question of existence and uniqueness results for the effective equation.

In §3, we examine the hyperbolic-parabolic equation in $]0,T[\times\Omega$:

$$(1.5) \qquad \rho\left(x, \frac{x}{\varepsilon}\right) \partial_t^2 u^\varepsilon - \operatorname{div}_x\left(k\left(x, \frac{x}{\varepsilon}\right) \operatorname{grad}_x u^\varepsilon\right) + \theta\left(\frac{x}{\varepsilon}, y, \frac{y}{\varepsilon}\right) \partial_t u^\varepsilon = f\left(t, x, \frac{x}{\varepsilon}, y, \frac{y}{\varepsilon}\right),$$

$$(1.6) \qquad u^\varepsilon\big|_{\partial\Omega_x} = 0 \quad \text{in} \quad]0,T[\times\mathcal{O},$$

$$(1.7) \qquad u^\varepsilon(0,x,y) = \alpha(x,y), \quad \sqrt{\rho\left(x, \frac{x}{\varepsilon}\right)}\, \partial_t u^\varepsilon(0,x,y) = \beta(x,y), \qquad (x,y) \in \Omega,$$

where $\Omega = \Omega_x \times \mathcal{O}$, Ω_x is an open bounded set in $\mathbb{R}^n$ with smooth boundary $\partial\Omega_x$, and $\rho \geq 0$ is allowed to vanish. First, averaging in the direction of propagation yields the damped wave equation

$$\tilde{\rho}(x)\,\partial_t^2 U - \operatorname{div}_x(\tilde{k}(x)\operatorname{grad}_x U) + \tilde{\theta}(y,\eta)\,\partial_t U = \tilde{f}(t,x,y,\eta), \tag{1.8}$$

$$U\,\big|_{\partial\Omega_x} = 0 \quad \text{in }]0,T[\times\mathcal{O}, \quad U\,\big|_{t=0} = \alpha, \quad \tilde{\rho}(x)\,\partial_t U\,\big|_{t=0} = \chi(x)\,\beta \quad \text{in } \Omega, \tag{1.9}$$

where $\tilde{\theta}(y,\eta) = \int_{Y_n} \theta(\zeta,y,\eta)\,d\zeta$, $\chi(x) = \int_{Y_n} \sqrt{\rho(x,\zeta)}\,d\zeta$. Next, the effective equation is obtained using the Dunford–Taylor's integral and the characterization of weak limits used in the preceding section. For this problem, there is no interaction between oscillations of the coefficients ρ and k with the damping coefficient θ. The nonlocal effects are due to oscillations of the damping and source terms. The main key results in §§2 and 3 leading to effective equations are valid only for equations where periodicity occurs in coefficients.

The last section is devoted to some similar problems involving the nonperiodic microstructure. We first analyze, for the sake of completeness, the propagation of a parametrized family of acoustic waves in a bounded domain Ω_x:

$$\partial_t^2 u^\varepsilon - a^\varepsilon(y)\,\Delta_x u^\varepsilon = 0, \quad u^\varepsilon\,|_{\partial\Omega_x} = 0, \quad u^\varepsilon\,|_{t=0} = \alpha, \quad \partial_t u^\varepsilon\,|_{t=0} = \beta. \tag{1.10}$$

We end the section by studying the damped wave equation

$$\begin{aligned} &\rho^\varepsilon(x)\,\partial_t^2 u^\varepsilon - \operatorname{div}_x(k^\varepsilon(x)\operatorname{grad}_x u^\varepsilon) + \theta^\varepsilon(y)\,\partial_t u^\varepsilon = 0, \\ &u^\varepsilon\,|_{\partial\Omega_x\times\mathcal{O}} = 0, \quad u^\varepsilon\,|_{t=0} = \alpha, \quad \partial_t u^\varepsilon\,|_{t=0} = \beta. \end{aligned} \tag{1.11}$$

We apply the same argument to show that there is no interaction when the damping coefficient does not depend on the direction x of the propagation.

A part of these results has been announced in [6].

2. A degenerate wave equation.

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^m$ $(m \geq 1)$ and $T > 0$. Setting $\Omega = \mathbb{R}^n \times \mathcal{O}$, $n \geq 1$, we consider the following Cauchy problem in $]0,T[\times\Omega$:

$$\rho^\varepsilon(x,y)\,\partial_t^2 u^\varepsilon - \operatorname{div}_x(k^\varepsilon(x,y)\operatorname{grad}_x u^\varepsilon) = f^\varepsilon(t,x,y), \tag{2.1}$$

$$u^\varepsilon(0,x,y) = \alpha(x,y), \quad \partial_t u^\varepsilon(0,x,y) = \beta(x,y), \quad (x,y) \in \Omega. \tag{2.2}$$

On the data ρ^ε, k^ε, f^ε, α, and β we make the following hypotheses. We suppose

$$\rho^\varepsilon(x,y) = \rho\left(\frac{x}{\varepsilon}, y, \frac{y}{\varepsilon}\right) \quad \text{for almost every } (x,y) \in \Omega, \tag{2.3}$$

where $\rho \in C(\overline{\mathcal{O}}; L^\infty_{\text{per}}(Y_n \times Y_m))$, that is, $\rho \in C(\overline{\mathcal{O}}; L^\infty(\mathbb{R}^n \times \mathbb{R}^m))$ and $\rho(\cdot,y,\cdot)$ is $Y_n \times Y_m$-periodic for every $y \in \mathcal{O}$, and

$$0 < \rho_- \leq \rho(\zeta,y,\eta) \leq \rho_+ \quad \text{a.e. in } Y_n \times \mathcal{O} \times Y_m; \tag{2.4}$$

the tensor k^ε is symmetric and satisfies

$$k^\varepsilon_{i,j}(x,y) = k_{i,j}\left(\frac{x}{\varepsilon}, y, \frac{y}{\varepsilon}\right) \quad \text{for almost every } (x,y) \in \Omega, \tag{2.5}$$

where, for $i,j=1$ to n, the components $k_{i,j}$ belongs to $C(\overline{\mathcal{O}}; L^\infty_{\text{per}}(Y_n \times Y_m))$ and

$$k_- \,|\xi|^2 \leq k(\zeta,y,\eta)\xi\cdot\xi \leq k_+ \,|\xi|^2 \quad \forall \xi \in \mathbb{R}^n, \tag{2.6}$$

for $y \in \mathcal{O}$, almost everywhere for $(\zeta,\eta) \in Y_n \times Y_m$, k_- and k_+ being two strictly positive real numbers. The source term f^ε is assumed to satisfy

$$\begin{aligned} &f^\varepsilon(t,x,y) = c^\varepsilon(x,y)\, f(t,x,y) \quad \text{for almost every } t \in]0,T[, (x,y) \in \Omega,\\ &c^\varepsilon(x,y) = c\left(\frac{x}{\varepsilon},\frac{y}{\varepsilon}\right), \qquad c_- \leq c^\varepsilon(x,y) \leq c_+ \quad \text{for almost every } (x,y) \in \Omega, \end{aligned} \tag{2.7}$$

where $c \in L^\infty_{\text{per}}(Y_n \times Y_m)$, and $f \in L^2(]0,T[\times\Omega)$. Concerning the initial data, we impose

$$\alpha \in L^2(\mathcal{O}; H^1(\mathbb{R}^n)), \qquad \beta \in L^2(\Omega). \tag{2.8}$$

It is well known that problem (2.1), (2.2) admits a unique solution; see Lions–Magenes [13] and Mizohata [14], for instance. Specifically the following assertions hold true:

(2.9) For each $\varepsilon > 0$, there is a unique solution to (2.1), (2.2) satisfying

$$u^\varepsilon \in C^0([0,T]; L^2(\mathcal{O}; H^1(\mathbb{R}^n))) \cap C^1([0,T]; L^2(\Omega)).$$

(2.10) The sequence $(u^\varepsilon)_{\varepsilon>0}$ is uniformly bounded in

$$L^\infty(0,T; L^2(\mathcal{O}; H^1(\mathbb{R}^n))) \cap W^{1,\infty}(0,T; L^2(\Omega)).$$

Let u be a weak limit, as $\varepsilon \to 0$, of a subsequence $(u^\varepsilon)_{\varepsilon>0}$ of solutions of problem (2.1), (2.2). We want to establish an effective equation satisfied by u. The way to construct that equation is like the process of reiterated homogenization. One first homogenizes with respect to ζ, while keeping y and η as fixed parameters, and next with respect to the direction η, which can be seen as a degenerate direction in (2.1).

2.1. Averaging with respect to ζ. A formal expansion for the solution u^ε of (2.1), (2.2) is obtained by the ansatz

$$u^\varepsilon(t,x,y) = U(t,x,\zeta,y,\eta) + \varepsilon\, U_1(t,x,\zeta,y,\eta) + \varepsilon^2\, U_2(t,x,\zeta,y,\eta) + \cdots \tag{2.11}$$

where $U(t,x,\zeta,y,\eta)$ and $U_j(t,x,\zeta,y,\eta)$ $j \geq 1$ are periodic in ζ and η. Following the usual procedure, we substitute (2.11) into the governing equation (2.1) and collect equal powers of ε. The coefficient of ε^{-2} gives

$$\operatorname{div}_\zeta \left(k(\zeta,y,\eta) \operatorname{grad}_\zeta U\right) = 0, \tag{2.12}$$

which implies that U does not depend on ζ. Hence $U = U(t,x,y,\eta)$. Taking (2.12) into account, the order ε^{-1} yields

$$\operatorname{div}_\zeta(k(\zeta,y,\eta)\,(\operatorname{grad}_x U + \operatorname{grad}_\zeta U_1)) = 0. \tag{2.13}$$

For the zero order, we have

$$\begin{aligned} \operatorname{div}_\zeta \left(k(\zeta,y,\eta) \operatorname{grad}_\zeta U_2\right) = &- c(\zeta,\eta)\, f(t,x,y) + \rho(\zeta,y,\eta)\,\partial_t^2 U\\ &- \operatorname{div}_\zeta(k(\zeta,y,\eta)\,(\operatorname{grad}_x U + \operatorname{grad}_\zeta U_1))\\ &- \operatorname{div}_\zeta(k(\zeta,y,\eta)\operatorname{grad}_x U_1). \end{aligned} \tag{2.14}$$

Thus, averaging (2.14) with respect to ζ, one obtains a wave equation for U:

$$\tilde{\rho}(y,\eta)\,\partial_t^2 U - \operatorname{div}_x(\tilde{k}(y,\eta)\operatorname{grad}_x U) = \tilde{c}(\eta)\, f(t,x,y) \tag{2.15}$$

for $t \in]0,T[$, $x \in \mathbb{R}^n$, $y \in \mathcal{O}$, and $\eta \in Y_m$, provided with the initial conditions

$$U(0,x,y,\eta) = \alpha(x,y), \quad \partial_t U(0,x,y,\eta) = \beta(x,y) \quad \text{in } \Omega \times Y_m. \tag{2.16}$$

Here $\tilde{c}(\eta) = \int_{Y_n} c(\zeta,\eta)\, d\zeta$, $\tilde{\rho}(y,\eta) = \int_{Y_n} \rho(\zeta,y,\eta)\, d\zeta$, and $\tilde{k}(y,\eta)$ is the homogenized tensor associated with $(k(x/\varepsilon,y,\eta))_{\varepsilon>0}$, for fixed y and η. Specifically $\tilde{k} = (\tilde{k}_{l,j})$, for $l,j = 1,\dots,n$ with

$$\tilde{k}_{l,j}(y,\eta) = \int_{Y_n} \left[k_{l,j}(\zeta,y,\eta) + \sum_{i=1}^{n} k_{i,j}(\zeta,y,\eta) \frac{\partial}{\partial \zeta_i} \chi_l(\zeta,y,\eta) \right] d\zeta. \tag{2.17}$$

The function $\chi_l = \chi_l(\zeta,y,\eta)$, with mean value $\langle \chi_l(\cdot,y,\eta)\rangle = \int_{Y_n} \chi_l(\zeta,y,\eta)\, d\zeta = 0$ is calculated explicitly by solving the cell problem

$$-\operatorname{div}_\zeta \left(k(\zeta,y,\eta)\,[e_l + \operatorname{grad}_\zeta \chi_l(\zeta,y,\eta)] \right) = 0, \qquad \zeta \in Y_n, \tag{2.18}$$

e_l being the lth vector of the canonical basis of $\mathbb{R}^n$. Notice that as $k(\zeta,y,\eta)$, $\tilde{k}(y,\eta)$ is also a symmetric tensor. To justify the above procedure, we shall pass to the limit in the variational formulation of (2.1), (2.2) with appropriate oscillatory test functions; see Bensoussan, Lions, and Papanicolaou [8], Murat [15], Sanchez-Palencia [18], and Tartar [19], and use the following convergence result due to Nguetseng [16].

THEOREM 2.1. *Let $(v^\varepsilon)_{\varepsilon>0}$ be a uniformly bounded sequence in $L^2_{\text{loc}}(\mathbb{R}^N)$. Then, there is a subsequence, still denoted by $(v^\varepsilon)_{\varepsilon>0}$, and a function V in the space $L^2_{\text{loc}}(\mathbb{R}^N;$ $L^2_{\text{per}}(Y_N))$ such that*

$$\int_{\mathbb{R}^N} v^\varepsilon(x)\,\psi\left(x,\frac{x}{\varepsilon}\right) dx \underset{\varepsilon\to 0}{\longrightarrow} \int_{\mathbb{R}^N\times Y_N} V(x,\zeta)\,\psi(x,\zeta)\, dx\, d\zeta \tag{2.19}$$

for any smooth function $\psi(x,\zeta)$ with compact support in x and periodic with respect to the variable ζ. If in addition $(\partial_{x_i} v^\varepsilon)_{\varepsilon>0}$ is uniformly bounded in $L^2_{\text{loc}}(\mathbb{R}^N)$, then

(i) *the function $V = V(x,\zeta)$ does not depend on the variable ζ_i;*

(ii) *there exists a function V_1, $V_1 = V_1(x,\zeta)$ with V_1 and $\partial_{\zeta_i} V_1$ in $L^2_{\text{loc}}(\mathbb{R}^N;$ $L^2_{\text{per}}(Y_N))$, such that*

$$\int_{\mathbb{R}^N} \partial_{x_i} v^\varepsilon(x)\,\psi\left(x,\frac{x}{\varepsilon}\right) dx \underset{\varepsilon\to 0}{\longrightarrow} \int_{\mathbb{R}^N\times Y_N} (\partial_{x_i} V + \partial_{\zeta_i} V_1)(x,\zeta)\,\psi(x,\zeta)\, dx\, d\zeta. \tag{2.20}$$

Remark 2.2. If $v^\varepsilon(x) = v(x,x/\varepsilon)$ in (2.19), where $v \in L^2_{\text{loc}}(\mathbb{R}^N; L^2_{\text{per}}(Y_N))$, then obviously $V(x,\zeta) = v(x,\zeta)$.

Remark 2.3. If $(v^\varepsilon)_{\varepsilon>0}$ is a bounded sequence in $L^2(\mathbb{R}^N)$ converging in the sense of equation (2.19) to $V = V(x,\zeta)$, then $(v^\varepsilon)_{\varepsilon>0}$ converges weakly in $L^2(\mathbb{R}^N)$ to $v(x) = \int_{Y_n} V(x,\zeta)\, d\zeta$.

Let us return to problem (2.1), (2.2). We are going to prove the following result.

THEOREM 2.4. *Let $(u^\varepsilon)_{\varepsilon>0}$ be the sequence of solutions of* (2.1), (2.2) *under assumptions* (2.3)–(2.8). *Then $(u^\varepsilon)_{\varepsilon>0}$ converges in the sense of* (2.19) *to a function*

$U = U(t,x,y,\eta)$ solution of the family of equations (2.15) *parametrized by* $(y,\eta) \in \mathcal{O} \times Y_m$ *with prescribed initial conditions* (2.16).

Proof. Let $\mathcal{L}$ denote the Laplace transform in t. We set for $p \in \mathbb{C}$ with $\Re e\, p > 0$

$$v^\varepsilon(p,x,y) = \mathcal{L}(u^\varepsilon(\cdot,x,y))(p), \qquad g(p,x,y) = \mathcal{L}(f(\cdot,x,y))(p) \quad \text{a.e. in } \Omega.$$

Taking the Laplace transform of (2.1), (2.2), one obtains

$$\begin{aligned} p^2\,\rho^\varepsilon(x,y)\,v^\varepsilon - \operatorname{div}_x(k^\varepsilon(x,y)\operatorname{grad}_x v^\varepsilon) &= c^\varepsilon(x,y)\,g(p,x,y) \\ &\quad + \rho^\varepsilon(x,y)\,(p\,\alpha(x,y) + \beta(x,y)). \end{aligned} \tag{2.21}$$

We now fix $p \in \mathbb{C}$ with $\Re e\, p > p_0 > 0$. The variational formulation of (2.21) reads

$$\begin{aligned} &\int_\Omega \left(p^2\,\rho^\varepsilon(x,y)\,v^\varepsilon\,\psi^\varepsilon + k^\varepsilon(x,y)\operatorname{grad}_x v^\varepsilon \cdot \operatorname{grad}_x \psi^\varepsilon\right) dx\,dy \\ &= \int_\Omega \{c^\varepsilon(x,y)\,g(p,x,y) + \rho^\varepsilon(x,y)\,(p\,\alpha(x,y) + \beta(x,y))\}\,\psi^\varepsilon\,dx\,dy \end{aligned} \tag{2.22}$$

for any test function ψ^ε. In the sequel φ stands for a function in $\mathcal{D}(\Omega)$. Choosing $\psi^\varepsilon(x,y) = \varepsilon\,\varphi(x,y)\,w(x/\varepsilon, y/\varepsilon)$ in (2.22), with w a smooth and $Y_n \times Y_m$-periodic function, we have

$$\lim_{\varepsilon\to 0} I^\varepsilon = 0, \quad \text{where } I^\varepsilon = \int_\Omega \varphi\, k^\varepsilon(x,y)\operatorname{grad}_x v^\varepsilon \cdot \operatorname{grad}_x w\left(\frac{x}{\varepsilon}, \frac{y}{\varepsilon}\right) dx\,dy.$$

On the other hand, Theorem 2.1 asserts that there is a subsequence (still denoted by $(v^\varepsilon)_{\varepsilon>0}$), $V = V(p,x,y,\eta)$, V in $L^2(\Omega; L^2_{\text{per}}(Y_m))$, $V_1 = V_1(p,x,\zeta,y,\eta)$, V_1, and $\partial_{\zeta_i} V_1$, in $L^2(\Omega; L^2_{\text{per}}(Y_n \times Y_m))$ for $i = 1$ to n, such that

$$\lim_{\varepsilon\to 0} I^\varepsilon = \int_{\Omega\times Y_n\times Y_m} \varphi\, k(\zeta,y,\eta)\,(\operatorname{grad}_x V + \operatorname{grad}_\zeta V_1) \cdot \operatorname{grad}_\zeta w\,dx\,dy\,d\zeta\,d\eta.$$

Hence

$$\operatorname{div}_\zeta\big(k(\zeta,y,\eta)\,(\operatorname{grad}_x V + \operatorname{grad}_\zeta V_1)\big) = 0 \quad \text{in the sense of distributions.} \tag{2.23}$$

Next we take $\psi^\varepsilon(x,y) = \varphi(x,y)\,w(y/\varepsilon)$ in (2.22), with w a smooth Y_m-periodic function, and pass to the limit as $\varepsilon \to 0$. In view of Theorem 2.1, we get

$$\begin{aligned} &p^2 \int_{\Omega\times Y_n\times Y_m} \rho(\zeta,y,\eta)\,\varphi\,V\,w(\eta)\,dx\,dy\,d\zeta\,d\eta \\ &+ \int_{\Omega\times Y_n\times Y_m} k(\zeta,y,\eta)\,\big(\operatorname{grad}_x V + \operatorname{grad}_\zeta V_1\big) \cdot \operatorname{grad}_x \varphi\,w(\eta)\,dx\,dy\,d\zeta\,d\eta \\ &= \int_{\Omega\times Y_n\times Y_m} \big(c(\zeta,\eta)\,g + \rho(\zeta,y,\eta)\,(p\,\alpha + \beta)\big)\,\varphi\,w(\eta)\,dx\,dy\,d\zeta\,d\eta. \end{aligned} \tag{2.24}$$

Relation (2.24) says that V and V_1 satisfy in $\mathcal{D}'(\Omega \times Y_m)$

$$\begin{aligned} &p^2\,\tilde{\rho}(y,\eta)\,V - \operatorname{div}_x\left(\int_{Y_n} k(\zeta,y,\eta)\,(\operatorname{grad}_x V + \operatorname{grad}_\zeta V_1)\,d\zeta\right) \\ &\quad = \tilde{\rho}(y,\eta)\,(p\,\alpha(x,y) + \beta(x,y)) + \tilde{c}(\eta)\,g(p,x,y). \end{aligned} \tag{2.25}$$

As in the classical framework of the homogenization theory, one can display from (2.23) V_1 in terms of V, and substituting it into (2.25) leads to

$$\begin{aligned} p^2\,\tilde\rho(y,\eta)\,V - \operatorname{div}_x(\tilde k(y,\eta)\operatorname{grad}_x V) \\ = \tilde\rho(y,\eta)\,(p\,\alpha(x,y)+\beta(x,y)) + \tilde c(\eta)\,g(p,x,y). \end{aligned} \tag{2.26}$$

Taking the inverse Laplace transform in (2.26), which is meaningful, there is a function $U = U(t,x,y,\eta)$ in $L^2(]0,T[\times\mathcal{O}\times Y_m; H^1(\mathbb{R}^n))$ solution of

$$\tilde\rho(y,\eta)\,\partial_t^2 U - \operatorname{div}_x(\tilde k(y,\eta)\operatorname{grad}_x U) = \tilde c(\eta)\,f(t,x,y), \tag{2.27}$$

$$U(0,x,y,\eta) = \alpha(x,y), \qquad \partial_t U(0,x,y,\eta) = \beta(x,y) \tag{2.28}$$

for $t\in]0,T[$, $x\in\mathbb{R}^n$, $y\in\mathcal{O}$, and $\eta\in Y_m$. Since the function U is uniquely determined by (2.27), (2.28), the whole sequence $(u^\varepsilon)_{\varepsilon>0}$ converges weakly in the sense of (2.19) to U. Theorem 2.4 is thereby proved.

2.2. Averaging with respect to η: effective equations. Let u be the zero-order moment of function U defined in Theorem 2.4, i.e.,

$$u(t,x,y) = \int_{Y_m} U(t,x,y,\eta)\,d\eta, \quad t\in]0,T[, \quad (x,y)\in\Omega. \tag{2.29}$$

Thanks to Theorem 2.1 and Remark 2.3, the sequence $(u^\varepsilon)_{\varepsilon>0}$ of solutions of (2.1), (2.2) converges as $\varepsilon\to 0$ to u weakly in $L^2(]0,T[\times\Omega)$. The homogenization of (2.1), (2.2) is now reduced to the averaging of (2.27) and (2.28) in Y_m. We shall then discuss such averaging subsequently. This problem enters in the framework of homogenizing parametrized families of wave equations. In order to be self-contained we summarize here the characterization of a weak* limit, obtained in [3], [4], which will be used throughout this paper.

Let (a^ε) and (b^ε) be two sequences in $L^\infty(\mathcal{O})$, not necessary in the periodic setting, satisfying

$$0 < a_- \le a^\varepsilon(y) \le a_+, \qquad 0 \le b^\varepsilon(y) \le b_+ \quad \text{a.e. in } \mathcal{O}.$$

Let $\Lambda =]a_-, a_+[$. We consider the sequence (ψ^ε) defined by

$$\psi^\varepsilon(y,z) = b^\varepsilon(y)\,(z - a^\varepsilon(y))^{-1} \quad \text{for all } z\in\mathbb{C}\setminus\overline{\Lambda} \quad \text{and for almost every } y\in\mathcal{O}.$$

Suppose

$$a^\varepsilon \overset{\star}{\rightharpoonup} \overline{a}, \quad \text{and} \quad b^\varepsilon \overset{\star}{\rightharpoonup} \overline{b} \quad \text{weakly* in } L^\infty(\mathcal{O}),$$

where $\overline{b}(y) \ge b_- > 0$ for almost every $y\in\mathcal{O}$. Define the coefficient $\overline{d}$ as

$$a^\varepsilon\, b^\varepsilon \overset{\star}{\rightharpoonup} \overline{b}\,\overline{d} \quad \text{weakly* in } L^\infty(\mathcal{O}).$$

The following lemma makes precise the weak* limit for a subsequence of (ψ^ε).

LEMMA 2.5. *There exists a parametrized family of nonnegative measures $d\sigma_y(\cdot)$, supported, for almost every $y\in\mathcal{O}$, in $\overline{\Lambda}$ such that for a subsequence*

$$\psi^\varepsilon(\cdot,z) \overset{\star}{\rightharpoonup} \psi(\cdot,z) \quad \textit{in } L^\infty(\mathcal{O}) \quad \textit{weak* for all } z\notin\overline{\Lambda}, \tag{2.30}$$

with

$$\psi(y,z)=\bar{b}(y)\left(z-\bar{d}(y)-\int_\Lambda (z-\lambda)^{-1}d\sigma_y(\lambda)\right)^{-1}. \tag{2.31}$$

We shall refer to $d\sigma_y$ as the parametrized family of measures associated with sequence $(a^\varepsilon, b^\varepsilon)$. The measure $d\sigma_y$ is defined by its moments $\int_\Lambda \lambda^k\, d\sigma_y(\lambda)$, which are explicitly given in terms of the limits of (a^ε), $(b^\varepsilon\,(a^\varepsilon)^k)$, $k \geq 0$. Let $\tau(y)=\int_\Lambda d\sigma_y(\lambda)$. One has $\tau(y)\geq 0$ and $\tau(y)\equiv 0$ if and only if the sequence $(a^\varepsilon(y))$ converges strongly. Let $d\nu_y$ be the family of Young measures associated with (ψ^ε). The support of $d\nu_y$ is contained in $[0,b_+]\times[a_-,a_+]$ and for all $z\in\mathbb{C}\setminus\overline{\Lambda}$, we have

$$\lim_{\varepsilon\to 0}\psi^\varepsilon(\cdot,z)=\langle d\nu_y(\mu,\lambda),\mu\,(z-\lambda)^{-1}\rangle.$$

We can state the following result.

LEMMA 2.6. *For almost every $y\in\mathcal{O}$, the measure $d\sigma_y$ is related to the measure $d\nu_y$ by*

$$\int_\Lambda d\sigma_y(\lambda)(z-\lambda)^{-1}=z-\bar{d}(y)-\bar{b}(y)\left(\int_{]0,b_+[\times\Lambda}\mu\,(z-\lambda)^{-1}\,d\nu_y(\mu,\lambda)\right)^{-1} \tag{2.32}$$

for all $z\in\mathbb{C}\setminus\overline{\Lambda}$. Furthermore, the family $d\sigma_y$ depends mesurably on y.

Proof. Relation (2.32) is established in [4], [5]. On the other hand, we know from Tartar [20] that the Young measure $d\nu_y$ depends measurably on $y\in\mathcal{O}$. Then, from (2.32), using Cauchy's formula we deduce that $y\mapsto d\sigma_y$ is also weakly measurable.

Let us now return to the averaging problem. According to whether the tensor $\tilde{k}$ defined by (2.17) is isotropic or not, the effective equations are obtained by different ways. We shall distinguish the following situations.

2.2.1. Case of macroscopic isotropy.

In this subsection, we suppose that the tensor $\tilde{k}(y,\eta)$ given by (2.17) is isotropic. For convenience, we denote also by $\tilde{k}(y,\eta)$ the scalar coefficient of the isotropic matrix. Consider the problem

$$\begin{aligned}&\tilde{\rho}(y,\eta)\,\partial_t^2U-\tilde{k}(y,\eta)\,\Delta_xU=\tilde{c}(\eta)\,f(t,x,y),\\ &U(0,x,y,\eta)=\alpha(x,y),\qquad \partial_tU(0,x,y,\eta)=\beta(x,y),\end{aligned} \tag{2.33}$$

for $t\in]0,T[$, $(x,y)\in\Omega$, $\eta\in Y_m$. Let $\widehat{U}(t,\xi,y,\eta)=\int_{\mathbb{R}}\exp(-2i\pi x\xi)\,U(t,x,y,\eta)\,dx$, the Fourier transform in x of U, and $V(p,\xi,y,\eta)=\mathcal{L}(\widehat{U}(\cdot,\xi,y,\eta))(p)$, the Laplace transform in time of $\widehat{U}$. Then for $p\in\mathbb{C}$ with $\Re e\,p>p_0>0$,

$$V(p,\xi,y,\eta)=\psi(p,\xi,y,\eta)\,(p\,\hat{\alpha}(\xi,y)+\hat{\beta}(\xi,y)+b(y,\eta)\,g(p,\xi,y)). \tag{2.34}$$

Here $\psi(p,\xi,y,\eta)=(p^2+4\pi^2|\xi|^2\,a(y,\eta))^{-1}$, $g(p,\xi,y)=\mathcal{L}(\hat{f}(\cdot,\xi,y))(p)$, $a(y,\eta)=\tilde{k}(y,\eta)/\tilde{\rho}(y,\eta)$ and $b(y,\eta)=\tilde{c}(\eta)/\tilde{\rho}(y,\eta)$. Owing to (2.4) and (2.6), one has

$$0<a_-\leq a(y,\eta)\leq a_+\quad\text{a.e. in }\mathcal{O}\times Y_m,\quad\text{with } a_-=\frac{k_-}{\rho_+},\quad a_+=\frac{k_+}{\rho_-}. \tag{2.35}$$

Let $v(p,\xi,y) = \int_{Y_m} V(p,\xi,y,\eta)\,d\eta$. From (2.34), we get

$$v(p,\xi,y) = \overline{\psi}\,(p\,\hat{\alpha}(\xi,y) + \hat{\beta}(\xi,y)) + \overline{b\psi}\,g(p,\xi,y),$$

where the overbar will denote the spatial averaging over the period Y_m. Notice that for fixed $\xi \in \mathbb{R}^n$ and $p \in \mathbb{C}$ with $\Re e\, p > p_0 > 0$, we have

$$\overline{\psi}(p,\xi,y) = \lim_{\varepsilon\to 0} \left(p^2 + 4\pi^2|\xi|^2\, a\left(y, \frac{y}{\varepsilon}\right)\right)^{-1},$$
$$\overline{b\psi}(p,\xi,y) = \lim_{\varepsilon\to 0} b\left(y, \frac{y}{\varepsilon}\right) \left(p^2 + 4\pi^2|\xi|^2\, a\left(y, \frac{y}{\varepsilon}\right)\right)^{-1} \quad \text{in } L^\infty(\mathcal{O}) \text{ weak*}.$$

Since we do not impose any condition on positiveness of coefficient $c^\varepsilon(x,y)$ in (2.7), we must be careful in applying Lemma 2.5, which holds only when c^ε does not change sign. Otherwise, we break up the coefficient c^ε in $c^\varepsilon(x,y) = c^\varepsilon_+(x,y) - c^\varepsilon_-(x,y)$, where c^ε_+ and c^ε_- are, respectively, the positive and negative parts of c^ε. As $\tilde{c}_\pm(\eta) \geq 0$, therefore,

$$\overline{b}_\pm(y) = \int_{Y_m} b_\pm(y,\eta)\,d\eta \geq \frac{\overline{c}_\pm}{\rho_+} > 0 \quad \text{with } \overline{c}_\pm = \int_{Y_m} \tilde{c}_\pm(\eta)\,d\eta > 0.$$

Let $\Lambda =]a_-, a_+[$, $\overline{a}(y) = \int_{Y_m} a(y,\eta)\,d\eta$. Lemma 2.5 asserts that there exist three families of nonnegative measures $d\sigma_y$, $d\sigma_y^+$ and $d\sigma_y^-$ associated, respectively, with sequences $(a(y,y/\varepsilon))_{\varepsilon>0}$, $(a(y,y/\varepsilon)b_+(y,y/\varepsilon))_{\varepsilon>0}$, and $(a(y,y/\varepsilon)b_-(y,\frac{y}{\varepsilon}))_{\varepsilon>0}$, parametrized by $y \in \mathcal{O}$ and supported in $\overline{\Lambda}$, such that

$$\text{(2.36)}\quad \overline{\psi}(p,\xi,y) = \left(p^2 + 4\pi^2|\xi|^2\,\overline{a}(y) - (4\pi^2|\xi|^2)^2 \int_\Lambda (p^2 + 4\pi^2|\xi|^2\,\lambda)^{-1}\,d\sigma_y(\lambda)\right)^{-1},$$

$$\text{(2.37)}\quad \begin{aligned}&\overline{b_\pm\,\psi}(p,\xi,y)\\ &= \overline{b}_\pm(y)\left(p^2 + 4\pi^2|\xi|^2\,\overline{d}_\pm(y) - (4\pi^2|\xi|^2)^2 \int_\Lambda (p^2 + 4\pi^2|\xi|^2\,\lambda)^{-1}\,d\sigma_y^\pm(\lambda)\right)^{-1},\end{aligned}$$

where

$$\overline{d}_\pm(y) = \frac{1}{\overline{b}_\pm(y)} \int_{Y_m} a(y,\eta)\,b_\pm(y,\eta)\,d\eta.$$

Let us now introduce functions $u_1 = u_1(t,x,y)$, $u_+ = u_+(t,x,y)$, and $u_- = u_-(t,x,y)$ such that

$$\text{(2.38)}\qquad \mathcal{L}(\hat{u}_1) = \overline{\psi}\,(p\,\hat{\alpha} + \hat{\beta}), \qquad \mathcal{L}(\hat{u}_\pm) = \overline{b_\pm\,\psi}\,g.$$

Using the inverse Laplace and Fourier transforms, it follows from (2.36), (2.37), and (2.38) that u_1, u_+ and u_- satisfy, respectively,

$$\text{(2.39)}\quad \begin{aligned}&\partial_t^2 u_1 - \overline{a}(y)\,\Delta_x u_1 - \int_0^t \int_{\mathbb{R}^n} M_n(t-s, x-z, y)\,\Delta_z^2 u_1(s,z,y)\,dz\,ds = 0,\\ &u_1\,|_{t=0} = \alpha, \qquad \partial_t u_1\,|_{t=0} = \beta,\end{aligned}$$

$$\begin{aligned}
&\partial_t^2 u_\pm - \overline{d}_\pm(y)\,\Delta_x u_2\\
&\qquad - \int_0^t \int_{\mathbb{R}^n} M_n^\pm(t-s,x-z,y)\,\Delta_x^2 u_\pm(s,z,y)\,dz\,ds = \overline{b}_\pm(y)\,f(t,x,y), \qquad (2.40)\\
&u_\pm\,|_{t=0} = 0, \qquad \partial_t u_\pm\,|_{t=0} = 0.
\end{aligned}$$

Here the kernels M_n, $M_n^\pm$ are given by

$$M_n(t,x,y) = \int_\Lambda d\sigma_y(\lambda)\,\mathcal{E}_n(t,x,\lambda), \qquad M_n^\pm(t,x,y) = \int_\Lambda d\sigma_y^\pm(\lambda)\,\mathcal{E}_n(t,x,\lambda),$$

where $\mathcal{E}_n(\cdot,\cdot,\lambda)$ is the elementary solution in $\mathbb{R}^n$ of the wave operator $\partial_t^2 - \lambda\,\Delta_x$, $\lambda \in \Lambda$.

We have thus proved the following.

THEOREM 2.7. *Let $(u^\varepsilon)_{\varepsilon>0}$ be the sequence of solutions of* (2.1), (2.2) *under assumptions* (2.3)–(2.8). *In the case of macroscopic isotropy, $(u^\varepsilon)_{\varepsilon>0}$ converges weakly in $L^2(]0,T[\times\Omega)$ to $u = u_1 + u_+ - u_-$, where u_1 and $u_\pm$ are characterized by* (2.39) *and* (2.40), *respectively.*

COROLLARY 2.8. *Assume that in* (2.3) *and* (2.6), *respectively, the functions ρ and c do not depend on the η-variable, that is, $\rho = \rho(\zeta,y)$, $c = c(\zeta)$. Hence, as $\varepsilon \to 0$, the sequence $(u^\varepsilon)_{\varepsilon>0}$ of solutions of* (2.1), (2.2) *converges weakly in $L^2(]0,T[\times\Omega)$ to u satisfying*

$$\begin{aligned}
&\tilde{\rho}(y)\,\partial_t^2 u - \overline{k}(y)\,\Delta_x u - \int_0^t \int_{\mathbb{R}^n} M_n(t-s,x-z,y)\,\Delta_x^2 u(s,z,y)\,dz\,ds = \tilde{c}\,f(t,x,y), \qquad (2.41)\\
&u_1\,|_{t=0} = \alpha, \qquad \partial_t u_1\,|_{t=0} = \beta.
\end{aligned}$$

Here $\tilde{\rho}(y) = \int_{Y_n} \rho(\zeta,y)\,d\zeta$, $\tilde{c} = \int_{Y_n} c(\zeta)\,d\zeta$, $\overline{k}(y) = \int_{Y_m} \tilde{k}(y,\eta)\,d\eta$ and $d\sigma_y$ represent the parametrized measure associated with $(\tilde{k}(y,y/\varepsilon))_{\varepsilon>0}$.

Remark 2.9. The one-dimensional case enters in the above pattern. The coefficient $\tilde{k}$ reduces to

$$\tilde{k}(y,\eta) = \int_0^1 \frac{d\zeta}{k(\zeta,y,\eta)}.$$

Remark 2.10. The Dunford–Taylor integral representation allows to extend the result for boundary value problem when the variable $x \in \Omega_x \subseteq \mathbb{R}^n$; see §4.

We establish now the existence and uniqueness result for equations of type (2.39), (2.40), or (2.41). Consider, for instance, (2.39). Introduce the auxiliary function w_1 defined on $]0,T[\,\times\Omega\times\Lambda$ by $w_1(t,x,y,\lambda) = \int_0^t \Delta_x u_1(s,x-\lambda(t-s),y)\,ds$. So (2.39) becomes the system

$$\begin{aligned}
&\partial_t^2 u_1 - \overline{a}(y)\,\Delta_x u_1 + \int_\Lambda \Delta_x w_1(t,x,y,\lambda)\,d\sigma_y(\lambda) = 0,\\
&\partial_t^2 w_1 - \lambda\,\Delta_x w_1 + \Delta_x u_1 = 0, \qquad (2.42)\\
&u_1\,|_{t=0} = \alpha, \quad \partial_t u_1\,|_{t=0} = \beta, \quad w_1\,|_{t=0} = 0, \quad \partial_t w_1\,|_{t=0} = 0.
\end{aligned}$$

We can write (2.42) as an abstract evolution problem in the form

$$\begin{aligned}
&U'' + AU = F, \qquad (2.43)\\
&U\,\Big|_{t=0} = U_0, \qquad U'\,\Big|_{t=0} = U_1.
\end{aligned}$$

The prime sign stands for the derivative with respect to time t, $U = (u, w)$, $U_0 = (\alpha_0, \beta_0)$, and $U_1 = (\alpha_1, \beta_1)$. The operator A is formally defined by

$$A \begin{pmatrix} u \\ w \end{pmatrix} = \begin{pmatrix} -\bar{a}(y)\, \Delta_x u + \int_\Lambda \Delta_x w(t,x,y,\lambda)\, d\sigma_y(\lambda) \\ \Delta_x u - \lambda\, \Delta_x w \end{pmatrix}. \tag{2.44}$$

Let $d\mu(y,\lambda) = dy\, d\sigma_y(\lambda)$. From Lemma 2.6 $d\mu$ is a nonnegative valued measure on $\mathcal{O} \times \Lambda$. Introduce then the Sobolev spaces $L^2_\mu(\mathcal{O} \times \Lambda) = L^2(\mathcal{O} \times \Lambda;\, d\mu(y,\lambda))$, $\mathbb{H} = L^2(\Omega) \times L^2_\mu(\mathcal{O} \times \Lambda; L^2(\mathbb{R}^n))$ and $\mathbb{V} = L^2(\mathcal{O}; H^1(\mathbb{R}^n)) \times L^2_\mu(\mathcal{O} \times \Lambda, H^1(\mathbb{R}^n))$. Relatively to the scalar product associated with a variational formulation of (2.43), one has

$$\begin{aligned}(AU, U) = &\int_\Omega \bar{a}(y)\, |\mathrm{grad}_x u|^2\, dx\, dy + \int_\Omega \int_\Lambda \lambda\, d\sigma_y(\lambda)\, |\mathrm{grad}_x w|^2\, dx\, dy \\ &- 2 \int_\Omega \int_\Lambda \mathrm{grad}_x u \cdot \mathrm{grad}_x w\, d\sigma_y(\lambda)\, dx\, dy.\end{aligned}$$

Let $\gamma > 0$. Since $2\, |\mathrm{grad}_x u \cdot \mathrm{grad}_x w| \le ((1+\gamma)/\lambda)\, |\mathrm{grad}_x u|^2 + (\lambda/(1+\gamma))\, |\mathrm{grad}_x w|^2$, it implies that

$$\begin{aligned}(AU, U) \ge &\int_\Omega |\mathrm{grad}_x u|^2 \left(\bar{a}(y) - (1+\gamma) \int_\Lambda \lambda^{-1} d\sigma_y(\lambda)\right) dx\, dy \\ &+ \frac{\gamma}{1+\gamma} \int_\Omega \int_\Lambda \lambda\, d\sigma_y(\lambda)\, |\mathrm{grad}_x w|^2\, dx\, dy.\end{aligned}$$

Taking (2.32) into account, one observes that

$$\bar{a}(y) - \int_\Lambda \lambda^{-1} d\sigma_y(\lambda) = \left(\lim_{\varepsilon \to 0} \frac{1}{a^\varepsilon(y)}\right)^{-1} = a_{-1}^{-1} \ge a_-.$$

It follows that

$$\bar{a}(y) - (1+\gamma) \int_\Lambda \lambda^{-1} d\sigma_y(\lambda) = -\gamma\, \bar{a}(y) + (1+\gamma)\, \bar{a}_{-1}^{-1}(y) \ge -\gamma\, a_+ + (1+\gamma)\, a_-.$$

Since $a_- > 0$, we can choose $\gamma > 0$ such that

$$\gamma_- = \min\left\{ -\gamma\, a_+ + (1+\gamma)\, a_-, \frac{\gamma}{1+\gamma}\, a_- \right\} > 0.$$

Thus

$$(AU, U) \ge \gamma_- \left(|\mathrm{grad}_x u|^2_{L^2(\Omega)} + |\mathrm{grad}_x w|^2_{L^2_\mu(\mathcal{O} \times \Lambda, L^2(\mathbb{R}^n))} \right). \tag{2.45}$$

Inequality (2.45) says that A is $\mathbb{V}$-elliptic. Hence, following Lions–Magenes [13, Thm. 8.1, p. 287], for U_1 in $\mathbb{H}$, U_0 in $\mathbb{V}$, and F in $L^2(0, T; \mathbb{H})$, there exists a unique function U satisfying (2.43) and such that

$$U \in L^2(0, T; \mathbb{V}) \quad \text{and} \quad U' \in L^2(0, T; \mathbb{H}).$$

Then we have established the following.

THEOREM 2.11. *Problem* (2.42) *admits a unique solution* $U = (u_1, w_1)$ *satisfying*

$$u_1 \in L^2(0,T; L^2(\mathcal{O}; H^1(\mathbb{R}^n))), \qquad w_1 \in L^2(0,T; L^2_\mu(\mathcal{O}\times\Lambda; H^1(\mathbb{R}^n))),$$
$$\partial_t u_1 \in L^2(0,T; L^2(\Omega)), \qquad \partial_t w_1 \in L^2(0,T; L^2_\mu(\mathcal{O}\times\Lambda; L^2(\Omega))).$$

Existence and uniqueness results for (2.40) and (2.41) can be handled in a similar manner. We conclude that the weak limit $u = u_1 + u_+ - u_-$ of sequence (u^ε) of solutions of (2.1), (2.2) is well defined by (2.39), (2.40), and (2.41).

2.2.2. Anisotropic case. The same basic approach as developed in [5] will be applied. First, we split up the solution U of (2.27), (2.28) into the superposition of solutions U_1 and U_2 of the subproblems

$$(2.46)\qquad \begin{aligned} &\tilde\rho(y,\eta)\,\partial_t^2 U_1 - \operatorname{div}_x(\tilde k(y,\eta)\operatorname{grad}_x U_1) = 0 \quad \text{in }]0,T[\times\Omega\times Y_m,\\ &U_1(0,x,y,\eta) = \alpha(x,y), \qquad \partial_t U_1(0,x,y,\eta) = \beta(x,y) \quad \text{in } \Omega\times Y_m, \end{aligned}$$

$$(2.47)\qquad \begin{aligned} &\tilde\rho(y,\eta)\,\partial_t^2 U_2 - \operatorname{div}_x(\tilde k(y,\eta)\operatorname{grad}_x U_2) = \tilde c(\eta)\, f(t,x,y) \quad \text{in }]0,T[\times\Omega\times Y_m,\\ &U_2(0,x,y,\eta) = 0, \qquad \partial_t U_2(0,x,y,\eta) = 0 \quad \text{in } \Omega\times Y_m. \end{aligned}$$

As before, we set $a(y,\eta) = \tilde k(y,\eta)/\tilde\rho(y,\eta)$ and $b(y,\eta) = \tilde c(\eta)/\tilde\rho(y,\eta)$. The tensor $a(y,\eta)$ satisfies

$$a_-\,|\xi|^2 \le a(y,\eta)\xi\cdot\xi \le a_+\,|\xi|^2 \quad \forall\xi\in\mathbb{R}^n, \quad \forall y\in\mathcal{O}, \quad \text{a.e. for } \eta\in Y_m,$$

with a_+ and a_- as in (2.35). To avoid complications in notation, we restrict ourselves to the case $0 \le \tilde c(\eta) \le c_+$. Then, according to the representation formula using Radon transform, we construct equations satisfied by the zero-order moment u_i of U_i, $i = 1, 2$, for which we just describe the procedure and refer to [5] for the details. We essentially use the isometry

$$(2.48)\qquad (-\partial_r^2)^{\frac{n-1}{4}}\mathcal{R}\colon L^2(\mathbb{R}^n) \longrightarrow L^2_o(\mathbb{R}\times S^{n-1}; |c_n|\,dr\,d\omega).$$

Here S^{n-1} denotes the unit sphere in $\mathbb{R}^n$, $c_n = (2i\pi)^{-n+1}/2$, $d\omega$ is the usual measure on S^{n-1}, $\mathcal{R}(\cdot)$ stands for the Radon transform, and the subscript o stands for even functions. Let, for $i = 1, 2$,

$$(2.49)\qquad \tilde u_i(t,r,\omega,y) = (-\partial_r^2)^{\frac{n-1}{4}}\,\mathcal{R}(u_i(t,\cdot,y))\,(r,\omega).$$

One can establish that $\tilde u_i$, $i = 1, 2$, satisfy, respectively,

$$(2.50)\qquad \begin{aligned} &\partial_t^2\tilde u_1 - \bar a(y)\,\omega\cdot\omega\,\partial_r^2\tilde u_1 + \int_\Lambda \partial_r^2\tilde w_1(t,r,\omega,y,\lambda)\,d\sigma^1_{y,\omega}(\lambda) = 0,\\ &\partial_t^2\tilde w_1 - \lambda\,\partial_r^2\tilde w_1 + \partial_r^2\tilde u_1 = 0, \end{aligned}$$

$$(2.51)\qquad \begin{aligned} &\partial_t^2\tilde u_2 - \bar a(y)\,\omega\cdot\omega\,\partial_r^2\tilde u_2 + \int_\Lambda \partial_r^2\tilde w_2(t,r,\omega,y,\lambda)\,d\sigma^2_{y,\omega}(\lambda) = \bar b(y)\,\tilde f(t,r,\omega),\\ &\partial_t^2\tilde w_2 - \lambda\,\partial_r^2\tilde w_2 + \partial_r^2\tilde u_2 = 0, \end{aligned}$$

where

$$\bar{b}(y)=\int_{Y_m} b(y,\eta)\,d\eta,\quad \bar{a}(y)=\int_{Y_m} a(y,\eta)\,d\eta \text{ and } \bar{d}(y)=\frac{1}{\bar{b}(y)}\int_{Y_m} a(y,\eta)\,b(y,\eta)\,d\eta.$$

For all fixed $\omega\in S^{n-1}$, $d\sigma^i_{y,\omega}$, $i=1,2$, are the measures associated, respectively, with the sequences $(a(y,y/\varepsilon)\omega\cdot\omega)_{\varepsilon>0}$ and $(b(y,y/\varepsilon)a(y,y/\varepsilon)\omega\cdot\omega)_{\varepsilon>0}$ in the sense of Lemma 2.5. Moreover, let $d\nu^1_y$ be the Young measure associated with the sequence $(a(y,y/\varepsilon))$. Then one has the relation

$$\begin{aligned}(2.52)\qquad &\langle d\sigma^1_{y,\omega}(\lambda),(z-\lambda)^{-1}\rangle\\ &=(\langle d\nu^1_y(T),(zI+T)\rangle\omega,\omega)-(\langle d\nu^1_y(T),(zI+T)^{-1}\rangle\omega,\omega)^{-1}.\end{aligned}$$

Here T stands for a matrix describing the oscillations of $a(y,y/\varepsilon)$. From (2.52), we deduce that, for any holomorphic function g in $\mathbb{C}$,

$$\langle d\Sigma^1_y(\cdot,\cdot),g(\cdot)\rangle\equiv\int_{S^{n-1}}\langle d\sigma^1_{y,\omega}(\cdot),g(\cdot)\rangle\,d\omega$$

is well defined and $d\Sigma^1_y(\cdot,\cdot)$ is a family of nonnegative measures, with support contained in $\overline{\Lambda}\times S^{n-1}$, parametrized by $y\in\mathcal{O}$. Hence, the inverse Radon transform gives the homogenized equation. Let $d\Sigma^2_y(\cdot,\cdot)$ be the corresponding measure associated with $d\sigma^2_{y,\omega}$.

We summarize this result in the following theorem.

THEOREM 2.12. *Assume that α, β, and f have compact supports in the direction x. Then, there exist two families of nonnegative measures $d\Sigma^i_y(\cdot,\cdot)$, $i=1,2$, parametrized by $y\in\mathcal{O}$, supported in $\overline{\Lambda}\times S^{n-1}$, such that the sequence of solutions $(u^\varepsilon)_{\varepsilon>0}$ of* (2.1), (2.2) *under assumptions* (2.3)–(2.8) *converges to $u=u_1+u_2$ weakly in $L^2(]0,T[\times\Omega)$. The functions u_1 and u_2 are solutions through auxiliary functions $w_1=w_1(t,x,y,\lambda)$, $w_2=w_2(t,x,y,\lambda)$, $(t,x,y)\in]0,T[\times\Omega$, $\lambda\in\Lambda$, of the following systems:*

(2.53)

$$\begin{aligned}&\partial_t^2u_1-\mathrm{div}_x(\bar{a}(y)\mathrm{grad}_x u_1)-c_n\int_{\Lambda\times S^{n-1}}d\Sigma^1_y(\lambda,\omega)\,(-\partial_r^2)^{\frac{n+1}{2}}\mathcal{R}(w_1)\,\big|_{r=x\cdot\omega}=0,\\ &\partial_t^2w_1-\lambda\,\Delta_x w_1+\Delta_x u_1=0,\quad t\in]0,T[,\quad (x,y)\in\mathbb{R}^n\times\mathcal{O},\quad \lambda\in\Lambda,\\ &u_1\,|_{t=0}=\alpha,\quad w_1\,|_{t=0}=0,\quad \partial_t u_1\,|_{t=0}=\beta,\quad \partial_t w_1\,|_{t=0}=0,\end{aligned}$$

$$\begin{aligned}(2.54)\qquad &\partial_t^2u_2-\mathrm{div}_x(\bar{d}(y)\mathrm{grad}_x u_2)\\ &\qquad -c_n\int_{\Lambda\times S^{n-1}}d\Sigma^2_y(\lambda,\omega)\,(-\partial_r^2)^{\frac{n+1}{2}}\mathcal{R}(w_2)\,\big|_{r=x\cdot\omega}=\bar{b}(y)\,f(t,x,y),\\ &\partial_t^2w_2-\lambda\,\Delta_x w_2+\Delta_x u_2=0,\quad t\in]0,T[,\quad (x,y)\in\mathbb{R}^n\times\mathcal{O},\quad \lambda\in\Lambda,\\ &u_2\,|_{t=0}=0,\quad w_2\,|_{t=0}=0,\quad \partial_t u_2\,|_{t=0}=0,\quad \partial_t w_2\,|_{t=0}=0.\end{aligned}$$

COROLLARY 2.13. *Assume that in* (2.3) *and* (2.6), *respectively, the functions* ρ *and* c *do not depend on the* η*-variable. Then, the weak limit* u *satisfies*

$$(2.55)\quad \begin{aligned} &\tilde{\rho}(y)\,\partial_t^2 u - \operatorname{div}_x(\overline{k}(y)\operatorname{grad}_x u) \\ &\qquad - c_n \int_{\Lambda\times S^{n-1}} d\Sigma_y(\lambda,\omega)\,(-\partial_r^2)^{\frac{n+1}{2}}\mathcal{R}(w)\,\big|_{r=x\cdot\omega} = \tilde{c}\, f(t,x,y), \\ &\tilde{\rho}(y)\,\partial_t^2 w - \lambda\,\Delta_x w + \Delta_x u = 0, \quad t \in]0,T[, \quad (x,y) \in \mathbb{R}^n \times \mathcal{O}, \quad \lambda \in \Lambda, \\ &u\,|_{t=0} = \alpha, \quad w\,|_{t=0} = 0, \quad \partial_t u\,|_{t=0} = \beta, \quad \partial_t w\,|_{t=0} = 0, \end{aligned}$$

where $\tilde{\rho}(y) = \int_{Y_n} \rho(\zeta, y)\,d\zeta$, $\tilde{c} = \int_{Y_n} c(\zeta)\,d\zeta$, $\overline{k}(y) = \int_{Y_m} \tilde{k}(y,\eta)\,d\eta$ *and* $d\Sigma_y$ *is the measure associated with* $(\tilde{k}(y, y/\varepsilon))$.

Finally, we state the following.

THEOREM 2.14. *Under hypotheses* (2.3)–(2.8), *problem* (2.53) (*respectively,* (2.54)) *admits a unique solution* $U = (u_1, w_1)$ (*respectively,* $U = (u_2, w_2)$) *satisfying for* $i = 1, 2$

$$\begin{aligned} &u_i \in L^2(0,T;L^2(\mathcal{O};H^1(\mathbb{R}^n))), &&w_i \in L^2(0,T;L^2_\mu(\mathcal{O}\times\Lambda;H^1(\mathbb{R}^n))), \\ &\partial_t u_i \in L^2(0,T;L^2(\Omega)), &&\partial_t w_i \in L^2(0,T;L^2_\mu(\mathcal{O}\times\Lambda;L^2(\Omega))). \end{aligned}$$

Thus, the weak limit $u = u_1 + u_2$ *of sequence* (u^ε) *of solutions of* (2.1), (2.2) *is well defined.*

Proof. Consider, for instance, system (2.53). Thanks to the isometry (2.48), it suffices to prove the existence and uniqueness for system (2.50). We can associate with system (2.50) an abstract formulation of type (2.43). Here the operator A is formally defined by

$$A\begin{pmatrix} u \\ w \end{pmatrix} = \begin{pmatrix} -\overline{a}(y)\omega\cdot\omega\,\partial_r^2 u + \int_\Lambda \partial_r^2 w(t,x,y,\lambda)\,d\sigma^1_{y,\omega}(\lambda) \\ \partial_r^2 u - \lambda\,\partial_r^2 w \end{pmatrix}.$$

Following the proof of Theorem 2.11, we introduce $d\mu(y,\omega,\lambda) = dy\,d\sigma^1_{y,\omega}(\lambda)\,d\omega$ and the Sobolev spaces $L^2_\mu(\mathcal{O}\times S^{n-1}\times\Lambda) = L^2(\mathcal{O}\times S^{n-1}\times\Lambda;\, d\mu(y,\omega,\lambda))$, $\widetilde{\mathbb{H}} = L^2(\widetilde{\Omega})\times L^2_\mu(\mathcal{O}\times S^{n-1}\times\Lambda;L^2(\mathbb{R}_+))$, and $\widetilde{\mathbb{V}} = L^2(\mathcal{O}\times S^{n-1}, H^1(\mathbb{R}_+))\times L^2_\mu(\mathcal{O}\times S^{n-1}\times\Lambda;H^1(\mathbb{R}_+))$, where $\widetilde{\Omega} = \mathbb{R}_+\times S^{n-1}\times\mathcal{O}$. One has

$$\begin{aligned} (AU,U) = &\int_{\widetilde{\Omega}} \overline{a}(y)\omega\cdot\omega\,|\partial_r u|^2\,dr\,dy\,d\omega + \int_{\widetilde{\Omega}}\int_\Lambda \lambda\,d\sigma^1_{y,\omega}(\lambda)\,|\partial_r w|^2\,d\sigma^1_{y,\omega}(\lambda)\,dr\,dy\,d\omega \\ &- 2\int_{\widetilde{\Omega}}\int_\Lambda \partial_r u\,\partial_r w\,d\sigma^1_{y,\omega}(\lambda)\,dr\,dy\,d\omega. \end{aligned}$$

For $\gamma > 0$, one deduces that

$$\begin{aligned} (AU,U) \geq &\int_{\widetilde{\Omega}} \left(\overline{a}(y)\omega\cdot\omega - (1+\gamma)\int_\Lambda \lambda^{-1}\,d\sigma^1_{y,\omega}(\lambda)\right)|\operatorname{grad}_x u|^2\,dr\,dy\,d\omega \\ &+ \frac{\gamma}{1+\gamma}\int_{\widetilde{\Omega}}\int_\Lambda \lambda\,d\sigma^1_{y,\omega}(\lambda)\,|\partial_r w|^2\,dr\,dy\,d\omega. \end{aligned}$$

From (2.52), one observes that

$$\overline{a}(y)\omega\cdot\omega-\int_{\Lambda}\lambda^{-1}\,d\sigma^{1}_{y,\omega}(\lambda)=\left(\lim_{\varepsilon\to 0}a\left(y,\frac{y}{\varepsilon}\right)^{-1}\omega\cdot\omega\right)^{-1}\geq a_{-}.$$

It follows that

$$\begin{aligned}\overline{a}(y)\omega\cdot\omega-(1+\gamma)\int_{\Lambda}\lambda^{-1}d\sigma^{1}_{y,\omega}(\lambda)&\geq-\gamma\,\overline{a}(y)\omega\cdot\omega+(1+\gamma)\,a_{-}\\&\geq-\gamma\,a_{+}+(1+\gamma)a_{-}.\end{aligned}$$

Thus, for some $\gamma_{-}>0$,

$$(AU,U)\geq\gamma_{-}\left(|\partial_{r}u|^{2}_{L^{2}(\widetilde{\Omega})}+|\partial_{r}w|^{2}_{L^{2}_{\mu}(\mathcal{O}\times S^{n-1}\times\Lambda,L^{2}(\mathbb{R}_{+}))}\right).$$

This completes the proof.

3. An hyperbolic-parabolic equation. Let $\mathcal{O}$ be an open subset of $\mathbb{R}^{m}$, $m\geq 1$, $T>0$, and let Ω_{x} be an open bounded set in $\mathbb{R}^{n}$ with smooth boundary $\partial\Omega_{x}$. We set $\Omega=\Omega_{x}\times\mathcal{O}$. We consider the following Dirichlet problem:

$$\rho^{\varepsilon}(x)\,\partial_{t}^{2}u^{\varepsilon}-\operatorname{div}_{x}(k^{\varepsilon}(x)\operatorname{grad}_{x}u^{\varepsilon})+\theta^{\varepsilon}(x,y)\,\partial_{t}u^{\varepsilon}=f^{\varepsilon}(t,x,y)\quad\text{in }]0,T[\times\Omega,\tag{3.1}$$

$$u^{\varepsilon}\big|_{\partial\Omega_{x}}=0\quad\text{in }]0,T[\times\mathcal{O},\tag{3.2}$$

$$u^{\varepsilon}\big|_{t=0}=\alpha\quad\text{in }\Omega,\tag{3.3}$$

$$\sqrt{\rho^{\varepsilon}}\,\partial_{t}u^{\varepsilon}\big|_{t=0}=\beta\quad\text{in }\Omega.\tag{3.4}$$

On the data ρ^{ε}, k^{ε}, θ^{ε}, f^{ε}, α, and β we assume that the following hypotheses hold. Let

$$\rho^{\varepsilon}(x)=\rho\left(x,\frac{x}{\varepsilon}\right)\quad\text{a.e. in }\Omega_{x},\tag{3.5}$$

with the function ρ belonging to $C(\overline{\Omega_{x}};L^{\infty}_{\text{per}}(Y_{n}))$, and satisfying

$$\begin{aligned}&0\leq\rho(x,\zeta)\leq\rho_{+}\quad\forall x\in\Omega_{x}\text{ for almost every }\zeta\in Y_{n},\\&\text{where we impose that }\tilde{\rho}(x)=\int_{Y_{n}}\rho(x,\zeta)\,d\zeta\geq\rho_{-}>0.\end{aligned}\tag{3.6}$$

The matrix k^{ε} is symmetric, with entries

$$k^{\varepsilon}_{i,j}(x)=k_{i,j}\left(x,\frac{x}{\varepsilon}\right)\quad\text{a.e. in }\Omega_{x},\quad i,j=1\text{ to }n,\tag{3.7}$$

where $k_{i,j}\in C(\overline{\Omega_{x}};L^{\infty}_{\text{per}}(Y_{n}))$, and satisfies

$$k_{-}\,|\xi|^{2}\leq k(x,\zeta)\xi\cdot\xi\leq k_{+}\,|\xi|^{2}\quad\forall\xi\in\mathbb{R}^{n}\quad\forall x\in\Omega_{x},\quad\text{a.e. for }\zeta\in\mathbb{R}^{n},\tag{3.8}$$

k_- and k_+ being two strictly positive real numbers. The damping coefficient is defined by

$$\theta^\varepsilon(x,y) = \theta\left(\frac{x}{\varepsilon}, y, \frac{y}{\varepsilon}\right) \quad \text{a.e. in } \Omega, \tag{3.9}$$

where $\theta \in C(\overline{\mathcal{O}}; L^\infty_{\text{per}}(Y_n \times Y_m))$, and

$$0 < \theta_- \le \theta(\zeta, y, \eta) \le \theta_+ \quad \forall y \in \mathcal{O} \quad \text{for almost every } (\zeta,\eta) \in Y_n \times Y_m. \tag{3.10}$$

In the sequel we set $\Lambda =]\theta_-, \theta_+[$. We take the source term f^ε as in (2.7) where we impose for the case of exposition

$$0 \le c\left(\frac{x}{\varepsilon}, \frac{y}{\varepsilon}\right) \le c_+. \tag{3.11}$$

On the initial data, we suppose

$$\alpha \in L^2(\mathcal{O}; H^1_0(\Omega_x)), \qquad \beta \in L^2(\Omega). \tag{3.12}$$

Observe that (3.4) gives a condition on $\partial_t u^\varepsilon \big|_{t=0}$ only on the set where $\rho^\varepsilon(x) \neq 0$. Hence (3.1) is an equation of hyperbolic-parabolic type. In the standard case where there is no dependence on the variable $y \in \mathcal{O}$, this class of equations has been considered in Vragov [21], and the homogenization is carried out in Bensoussan–Lions–Papanicolaou [8]. Under assumptions (3.5)–(3.12) problem (3.1)–(3.4) admits a unique solution in a weak sense; see [13], [14]. More precisely, we have

(3.13) For each $\varepsilon > 0$, there is a unique solution to (3.1)–(3.4) satisfying

$$u^\varepsilon \in L^\infty(0,T; L^2(\mathcal{O}; H^1_0(\Omega_x))) \cap C^0([0,T]; L^2(\Omega)).$$

(3.14) The sequence $(u^\varepsilon)_{\varepsilon>0}$ is uniformly bounded in

$$L^\infty(0,T; L^2(\mathcal{O}; H^1_0(\Omega_x))) \cap H^1(0,T; L^2(\Omega)).$$

Furthermore

$$\sqrt{\rho^\varepsilon}\, \partial_t u^\varepsilon \quad \text{is bounded in } L^2(]0,T[\times\Omega). \tag{3.15}$$

The proof of (3.13)–(3.15) is standard. It relies on the following a priori identity of energy

$$\begin{aligned} &\int_\Omega \rho^\varepsilon(x)|\partial_t u^\varepsilon|^2\,dx\,dy + \int_\Omega k^\varepsilon(x)\operatorname{grad}_x u^\varepsilon \cdot \operatorname{grad}_x u^\varepsilon\,dx\,dy \\ &\quad + \int_0^t \int_\Omega \theta^\varepsilon(x,y)|\partial_t u^\varepsilon|^2\,dx\,dy\,ds \\ &= \int_\Omega |\beta|^2\,dx\,dy + \int_\Omega k^\varepsilon(x)\operatorname{grad}_x \alpha \cdot \operatorname{grad}_x \alpha\,dx\,dy \\ &\quad + \int_0^t \int_\Omega f^\varepsilon(s,x,y)\,\partial_t u^\varepsilon\,dx\,dy\,ds \end{aligned} \tag{3.16}$$

for all $t \in (0,T)$.

Remark 3.1. Since (u^ε) belongs to $C^0([0,T];L^2(\Omega))$, (3.3) has a meaning. From (3.5), (3.6), (3.13), and (3.14) it follows that $\rho^\varepsilon\,\partial_t u^\varepsilon$ belongs to $L^2(0,T;L^2(\Omega))$ and $\partial_t(\rho^\varepsilon\,\partial_t u^\varepsilon)$ belongs to $L^2(0,T;L^2(\mathcal{O};H^{-1}(\Omega_x)))$. Thus

$$\rho^\varepsilon\,\partial_t u^\varepsilon \quad \text{belongs to } C^0([0,T];H^{-1}(\Omega)). \tag{3.17}$$

We know from [13, Lem. 8.1, p. 297] that

$$L^\infty(0,T;L^2(\Omega))\cap C^0([0,T];H^{-1}(\Omega))\subset C_s([0,T];L^2(\Omega)),$$

where $C_s([0,T];L^2(\Omega))$ denotes the space of functions v defined in $[0,T]$ with values in $L^2(\Omega)$, such that for any w in $L^2(\Omega)$ the function $t\longmapsto \int_\Omega v(t)\,w\,dx\,dy$ is continuous on $[0,T]$. Therefore $\rho^\varepsilon\,\partial_t u^\varepsilon$ belongs to $C_s([0,T];L^2(\Omega))$, and consequently the function

$$t\longmapsto \int_\Omega (\sqrt{\rho^\varepsilon}\,w)\,(\sqrt{\rho^\varepsilon}\,\partial_t u^\varepsilon)\,dx\,dy$$

is continuous on $[0,T]$ for any w such that $\sqrt{\rho^\varepsilon}\,w\in L^2(\Omega)$. This implies that the initial condition (3.4) has a meaning. Obviously the condition $\rho^\varepsilon\,\partial_t u^\varepsilon\,\big|_{t=0}=\sqrt{\rho^\varepsilon}\,\beta$ is also well defined.

In order to derive effective equation describing the oscillatory behavior of (u^ε), we shall proceed as in §2.

3.1. Averaging with respect to ζ. Taking $y\in\mathcal{O}$ and $\eta\in Y_m$ as fixed parameters, we define

$$\tilde c(\eta)=\int_{Y_n} c(\zeta,\eta)\,d\zeta,\ \tilde\theta(y,\eta)=\int_{Y_n}\theta(\zeta,y,\eta)\,d\zeta,\quad\text{and}\quad \chi(x)=\int_{Y_n}\sqrt{\rho(x,\zeta)}\,d\zeta. \tag{3.18}$$

Let also $\tilde k(x)$ be the homogenized matrix associated with the sequence $(k(x,x/\varepsilon))_{\varepsilon>0}$.

We now state and prove the following result.

THEOREM 3.2. *Let $(u^\varepsilon)_{\varepsilon>0}$ be the sequence of solutions of* (3.1)–(3.4), *under assumptions* (3.5)–(3.12). *Then, there is a unique function $U=U(t,x,y,\eta)$ in $L^2(]0,T[\times\mathcal{O}\times Y_m;H_0^1(\Omega_x))$ such that $(u^\varepsilon)_{\varepsilon>0}$ converges, as $\varepsilon\to 0$, to U in the sense of* (2.19). *For almost every $(y,\eta)\in\mathcal{O}\times Y_m$, the function $U(\cdot,\cdot,y,\eta)$ is the solution of*

$$\tilde\rho(x)\,\partial_t^2 U-\operatorname{div}_x(\tilde k(x)\operatorname{grad}_x U)+\tilde\theta(y,\eta)\,\partial_t U=\tilde c(\eta)\,f(t,x,y) \tag{3.19}$$

for $t\in]0,T[$, $x\in\Omega_x$,

$$U\,\big|_{t=0}=\alpha\quad \textit{in}\ \Omega_x, \tag{3.20}$$

$$\tilde\rho(x)\,\partial_t U\,\big|_{t=0}=\chi(x)\,\beta\quad \textit{in}\ \Omega_x. \tag{3.21}$$

Remark 3.3. Assume that in (3.5) the function ρ is independent of ζ and $\rho\not\equiv 0$. Then, it follows from (3.21) that $\partial_t U\,\big|_{t=0}$ is given over the whole domain Ω_x. There is some sort of increase of the initial data comparing with (3.4); see [8].

Proof. It is very close to that of Theorem 2.4. Without going into further details, let us outline the procedure. Let $v^\varepsilon(p,x,y)=\mathcal{L}(u^\varepsilon(\cdot,x,y))(p)$, $g(p,x,y)=$

$\mathcal{L}(f(\cdot,x,y))(p)$ and fix $p \in \mathbb{C}$ with $\Re e p > p_0$ for some real $p_0 > 0$. Therefore, by Laplace transform, (3.1)–(3.4) becomes

$$\begin{aligned} &p^2\,\rho^\varepsilon(x)\,v^\varepsilon - \operatorname{div}_x(k^\varepsilon(x)\operatorname{grad}_x v^\varepsilon) + p\,\theta^\varepsilon(x,y)\,v^\varepsilon \\ &= c^\varepsilon(x,y)\,g(p,x,y) + (p\,\rho^\varepsilon(x) + \theta^\varepsilon(x,y))\,\alpha(x,y) + \sqrt{\rho^\varepsilon(x)}\,\beta(x,y), \end{aligned} \tag{3.22}$$

$$v^\varepsilon\,\big|_{\partial\Omega_x} = 0. \tag{3.23}$$

Consider the variational formulation of (3.22), (3.23):

$$\begin{aligned} &\int_\Omega \left((p^2\,\rho^\varepsilon(x) + p\,\theta^\varepsilon(x,y))\,v^\varepsilon\,\psi^\varepsilon + k^\varepsilon(x)\operatorname{grad}_x v^\varepsilon\cdot\operatorname{grad}_x\psi^\varepsilon\right)dx\,dy \\ &= \int_\Omega \left\{(p\,\rho^\varepsilon(x) + \theta^\varepsilon(x,y))\,\alpha(x,y) + \sqrt{\rho^\varepsilon(x)}\,\beta(x,y)\right\}\psi^\varepsilon\,dx\,dy \\ &\quad + \int_\Omega c^\varepsilon(x,y)\,g(p,x,y)\,\psi^\varepsilon\,dx\,dy \end{aligned} \tag{3.24}$$

for any test function ψ^ε. Arguing like before, we take successively in (3.24)

$$\psi^\varepsilon(x,y) = \varepsilon\,\varphi(x,y)\,w\left(\frac{x}{\varepsilon},\frac{y}{\varepsilon}\right) \quad\text{and}\quad \psi^\varepsilon(x,y) = \varphi(x,y)\,w\left(\frac{y}{\varepsilon}\right)$$

with $\varphi \in \mathcal{D}(\Omega)$, w a smooth $Y_n\times Y_m$-periodic function and send $\varepsilon \to 0$. Thanks to Theorem 2.1, there exists a subsequence (still denoted by $(v^\varepsilon)_{\varepsilon>0}$), and $V = V(p,x,y,\eta)$, $V_1 = V_1(p,x,\zeta,y,\eta)$, V in $L^2(\Omega; L^2_{\text{per}}(Y_m))$, V_1 and $\partial_{\zeta_i}V_1$ in $L^2(\Omega; L^2_{\text{per}}(Y_n\times Y_m))$ such that

$$\operatorname{div}_\zeta(k(x,\zeta)\,(\operatorname{grad}_x V + \operatorname{grad}_\zeta V_1)) = 0, \tag{3.25}$$

$$\begin{aligned} &p^2\,\tilde\rho(x)\,V - \operatorname{div}_x\left(\int_{Y_n} k(x,\zeta)\,(\operatorname{grad}_x V + \operatorname{grad}_\zeta V_1)\right)d\zeta + p\,\tilde\theta(y,\eta)\,V \\ &= (p\,\tilde\rho(x) + \tilde\theta(y,\eta))\,\alpha(x,y) + \chi(x)\,\beta(x,y) + \tilde c(\eta)\,g(p,x,y). \end{aligned} \tag{3.26}$$

in the sense of distributions. Displaying from (3.25) V_1 in terms of V and introducing the homogenized tensor $\tilde k(x)$ of $(k(x,x/\varepsilon))_{\varepsilon>0}$ gives

$$\begin{aligned} &p^2\,\tilde\rho(x)\,V - \operatorname{div}_x(\tilde k(x)\operatorname{grad}_x V) + p\,\tilde\theta(y,\eta)\,V \\ &= (p\,\tilde\rho(x) + \tilde\theta(y,\eta))\,\alpha(x,y) + \chi(x)\,\beta(x,y) + \tilde c(\eta)\,g(p,x,y). \end{aligned} \tag{3.27}$$

Taking the inverse Laplace transform of (3.27), one deduces that there is a function $U = U(t,x,y,\eta)$ in $L^2(]0,T[\times\mathcal{O}\times Y_m; H^1_0(\Omega_x))$ such that (3.19), (3.20), and (3.21) hold. Moreover, the function U is uniquely determined by (3.19)–(3.21). Hence the whole sequence $(u^\varepsilon)_{\varepsilon>0}$ converges weakly in the sense of (2.19) to U. The proof of Theorem 3.2 is then complete.

3.2. Averaging with respect to η. Nonlocal effective equations. Let u be the zero-order moment of the function U defined in Theorem 3.2

$$u(t,x,y) = \int_{Y_m} U(t,x,y,\eta)\,d\eta, \quad t>0, \quad (x,y)\in\Omega. \tag{3.28}$$

Clearly, the sequence $(u^\varepsilon)_{\varepsilon>0}$ of solutions of (3.1)–(3.4) converges as $\varepsilon \to 0$ to u weakly in $L^2(]0,T[\times\Omega)$. In order to get an equation for u, it remains to average (3.19) with respect to $\eta \in Y_m$. Let $V(p,x,y,\eta) = \mathcal{L}(U(\cdot,x,y,\eta))(p)$, the Laplace transform in time t of U. Then for $p \in \mathbb{C}$ with $\Re e p > 0$, V satisfies (3.27). For fixed $p \in \mathbb{C}$ with $\Re e p > p_0$, introduce the unbounded operator A_p with domain $D(A_p) = \{v \in H_0^1(\Omega_x)\,;\ A_p v \in L^2(\Omega_x)\}$ defined by

$$A_p v = p^2\,\tilde{\rho}(x)\,v - \mathrm{div}_x(\tilde{k}(x)\,\mathrm{grad}_x v). \tag{3.29}$$

For almost every $(y,\eta) \in \mathcal{O}\times Y_m$, $(A_p + p\,\tilde{\theta}(y,\eta))$ is a one-to-one operator from $D(A_p)$ onto $L^2(\Omega_x)$. Then, from (3.27) we deduce

$$V(p,\cdot,y,\eta) = \left(A_p + p\,\tilde{\theta}(y,\eta)\right)^{-1}\left(h(p,x,y,\eta)\right) \tag{3.30}$$

with $h(p,x,y,\eta) = (p\,\tilde{\rho}(x) + \tilde{\theta}(y,\eta))\,\alpha(x,y) + \chi(x)\,\beta(x,y) + \tilde{c}(\eta)\,g(p,x,y)$. According to Dunford–Taylor's integral (see Kato [12]) we can write

$$\left(A_p + p\,\tilde{\theta}(y,\eta)\right)^{-1} = \frac{-1}{2\,i\,\pi}\int_{\Gamma_p}\left(z - p\,\tilde{\theta}(y,\eta)\right)^{-1}(A_p + z)^{-1}\,dz. \tag{3.31}$$

Here

$$\Gamma_p = \left\{z \in \mathbb{C}\,;\, \frac{z}{p} \in \tilde{\Gamma}_p\right\}$$

with $\tilde{\Gamma}_p$ a closed curve in the right half plane $\Re e\, z > 0$ containing the real interval $\overline{\Lambda}$. Thanks to Fubini's lemma and Lemma 2.5, averaging in Y_m of (3.31) is easily obtained. There exist two parametrized families of nonnegative measures $d\sigma_y^1$ and $d\sigma_y^2$ associated, respectively, with the sequences $(\tilde{\theta}(y,y/\varepsilon))_{\varepsilon>0}$ and $(\tilde{c}(y/\varepsilon)\,\tilde{\theta}(y,y/\varepsilon))_{\varepsilon>0}$, with support contained in $\overline{\Lambda}$ such that for all $z \in \Gamma_p$ and almost every $y \in \mathcal{O}$,

$$\int_{Y_m}(z - p\,\tilde{\theta}(y,\eta))^{-1}\,d\eta = \left(z - p\,\overline{\theta}(y) - p^2\int_\Lambda (z - p\,\lambda)^{-1}, d\sigma_y^1(\lambda)\right)^{-1}, \tag{3.32}$$

$$\int_{Y_m}\tilde{c}(\eta)\,(z - p\,\tilde{\theta}(y,\eta))^{-1}\,d\eta = \overline{c}\left(z - p\,\overline{d}(y) - p^2\int_\Lambda (z - p\,\lambda)^{-1}\,d\sigma_y^2(\lambda)\right)^{-1}, \tag{3.33}$$

where $\overline{\theta}(y) = \int_{Y_m}\tilde{\theta}(y,\eta)\,d\eta$, $\overline{c} = \int_{Y_m}\tilde{c}(\eta)\,d\eta$, and $\overline{d}(y) = \frac{1}{\overline{c}}\int_{Y_m}\tilde{\theta}(y,\eta)\,\tilde{c}(\eta)\,d\eta$. Introduce for $i = 1,2$, the functions $V_i = V_i(p,x,y,\eta)$ such that

$$\begin{aligned} V_1(p,x,y,\eta) &= (A_p + p\,\tilde{\theta}(y,\eta))^{-1}\left((p\,\tilde{\rho}(x) + \tilde{\theta}(y,\eta))\,\alpha(x,y) + \chi(x)\,\beta(x,y)\right),\\ V_2(p,x,y,\eta) &= (A_p + p\,\tilde{\theta}(y,\eta))^{-1}\left(\tilde{c}(\eta)\,g(p,x,y)\right). \end{aligned} \tag{3.34}$$

First, it follows from Dunford–Taylor's integral that

$$\begin{aligned} &\int_{Y_m}(A_p + p\,\tilde{\theta}(y,\eta))^{-1}(h(p,\cdot,y))\,d\eta\\ &\quad = \left(A_p + p\,\overline{\theta}(y) - p^2\int_\Lambda (A_p + p\,\lambda)^{-1}\,d\sigma_y^2(\lambda)\right)^{-1}(h(p,\cdot,y)) \end{aligned} \tag{3.35}$$

for any function $h(p,\cdot,y)$ in $L^2(\Omega_x)$. Using the relation

$$p\,\tilde\theta(y,\eta)\,(z-p\,\tilde\theta(y,\eta))^{-1} = -1 + z\,(z-p\,\tilde\theta(y,\eta))^{-1}, \tag{3.36}$$

we also obtain

$$\begin{aligned}&\int_{Y_m}\tilde\theta(y,\eta)\,(A_p+p\,\tilde\theta(y,\eta))^{-1}(h(p,\cdot,y))\,d\eta\\ &\quad= \left(A_p+p\,\bar\theta(y)-p^2\int_\Lambda (A_p+p\,\lambda)^{-1}\,d\sigma_y^2(\lambda)\right)^{-1}(h_*(p,\cdot,y))\end{aligned} \tag{3.37}$$

with $h_*(p,\cdot,y) = (\bar\theta(y) - p\int_\Lambda (A_p+p\,\lambda)^{-1}\,d\sigma_y^1(\lambda))(h(p,\cdot,y))$. Similarly, we deduce from (3.33) that

$$\begin{aligned}&\int_{Y_m}(A_p+p\,\tilde\theta(y,\eta))^{-1}(h(p,\cdot,y)\,\tilde b(y,\eta))\,d\eta\\ &\quad= \bar c\left(A_p+p\,\bar d(y)-p^2\int_\Lambda (A_p+p\,\lambda)^{-1}\,d\sigma_y^2(\lambda)\right)^{-1}(h(p,\cdot,y)).\end{aligned} \tag{3.38}$$

Now let $v(t,x,y)=\int_{Y_m} V(p,x,y,\eta)\,d\eta$ and $v_i(t,x,y)=\int_{Y_m} V_i(p,x,y,\eta)\,d\eta$ for $i=1,2$. The functions $v_i(p,\cdot,y)$, $i=1,2$, belong to $H_0^1(\Omega_x)$ and $v=v_1+v_2$. Also let introduce v_1^* and v_2^* such that

$$v_1^* = (A_p+p\,\lambda)^{-1}(p\,v_1+\alpha), \qquad v_2^* = (A_p+p\,\lambda)^{-1}(p\,v_2).$$

Therefore, from (3.35), (3.37), and (3.38) one easily deduces that (v_1,v_2,v_1^*,v_2^*) satisfy the following system of equations:

$$\begin{aligned}&A_p\,v_1+p\,\bar\theta(y)\,v_1-p\int_\Lambda v_1^*\,d\sigma_y^1(\lambda) = p\,(\tilde\rho(\cdot)+\bar\theta(y))\,\alpha(\cdot,y)+\chi(\cdot)\,\beta(\cdot,y),\\ &A_p\,v_1^*+p\,\lambda\,v_1^*-p\,v_1=\alpha,\\ &A_p\,v_2+p\,\bar d(y)\,v_2-p\int_\Lambda v_2^*\,d\sigma_y^2(\lambda)=\bar c\,g(p,\cdot,y),\\ &A_p\,v_2^*+p\,\lambda\,v_2^*-p\,v_2=0.\end{aligned}$$

The inverse Laplace transform, with the notation $u_i=\mathcal{L}^{-1}(v_i)$, $w_i=\mathcal{L}^{-1}(v_i^*)$, gives

$$\begin{aligned}&\tilde\rho(x)\,\partial_t^2u_1-\operatorname{div}_x(\tilde k(x)\operatorname{grad}_x u_1)+\bar\theta(y)\,\partial_t u_1-\int_\Lambda \partial_t w_1\,d\sigma_y^1(\lambda)=0,\\ &\tilde\rho(x)\,\partial_t^2w_1-\operatorname{div}_x(\tilde k(x)\operatorname{grad}_x w_1)+\lambda\,\partial_t w_1-\partial_t u_1=0,\\ &u_1\,\big|_{\partial\Omega_x\times\mathcal{O}}=0,\\ &u_1\,\big|_{t=0}=\alpha,\qquad \tilde\rho\,\partial_t u_1\,\big|_{t=0}=\chi\cdot\beta\quad\text{in }\Omega,\\ &w_1\,\big|_{t=0}=0,\qquad \partial_t w_1\,\big|_{t=0}=0\quad\text{in }\Omega,\end{aligned} \tag{3.39}$$

$$\begin{aligned}&\tilde\rho(x)\,\partial_t^2u_2-\operatorname{div}_x(\tilde k(x)\operatorname{grad}_x u_2)+\bar d(y)\,\partial_t u_2-\int_\Lambda \partial_t w_2\,d\sigma_y^2(\lambda)=0,\\ &\tilde\rho(x)\,\partial_t^2w_2-\operatorname{div}_x(\tilde k(x)\operatorname{grad}_x w_2)+\lambda\,\partial_t w_2-\partial_t u_2=\bar c\,f(t,x,y),\\ &u_2\,\big|_{\partial\Omega_x\times\mathcal{O}}=0,\\ &u_2\,\big|_{t=0}=0,\qquad \partial_t u_2\,\big|_{t=0}=0\quad\text{in }\Omega,\\ &w_2\,\big|_{t=0}=0,\qquad \partial_t w_2\,\big|_{t=0}=0\quad\text{in }\Omega.\end{aligned} \tag{3.40}$$

Thus we obtain the following.

THEOREM 3.4. *Let* $(u^\varepsilon)_{\varepsilon>0}$ *be the sequence of solutions of* (3.1)–(3.4) *under assumptions* (3.5)–(3.9). *Then,* $(u^\varepsilon)_{\varepsilon>0}$ *converges weakly in* $L^2(]0,T[\times\Omega)$ *to* $u = u_1+u_2$, *where* u_1 *and* u_2 *are, respectively, defined by* (3.39) *and* (3.40).

Remark 3.5. If in (3.11) the function c does not depend on the η-variable, then the corresponding sequence of solutions of (3.1)–(3.4) converges weakly in $L^2(]0,T[\times\Omega)$ to u solution of the following system:

$$\tag{3.41}\begin{aligned}
&\tilde\rho(x)\,\partial_t^2 u - \operatorname{div}_x(\tilde k(x)\operatorname{grad}_x u) + \overline{\theta}(y)\,\partial_t u - \int_\Lambda \partial_t w\, d\sigma_y^1(\lambda) = \tilde c\, f(t,x,y),\\
&\tilde\rho(x)\,\partial_t^2 w - \operatorname{div}_x(\tilde k(x)\operatorname{grad}_x w) + \lambda\,\partial_t w - \partial_t u = 0,\\
&u\,\Big|_{\partial\Omega_x\times\mathcal{O}} = 0,\\
&u\,\Big|_{t=0} = \alpha, \qquad \tilde\rho\,\partial_t u\,\Big|_{t=0} = \chi\beta \quad \text{in } \Omega,\\
&w\,\Big|_{t=0} = 0, \qquad \partial_t w\,\Big|_{t=0} = 0 \quad \text{in } \Omega.
\end{aligned}$$

Remark 3.6. Existence and uniqueness results for (3.39)–(3.41), may be established by the same technique as in Theorem 2.11.

Remark 3.7. If the function c^ε changes sign, we proceed as in §2.

4. Parametrized families in the nonperiodic microstructure. Let $T > 0$ and $\Omega = \Omega_x \times \mathcal{O}$, where Ω_x is an open bounded set in $\mathbb{R}^n$ with smooth boundary $\partial\Omega_x$ and $\mathcal{O}$ is a domain in $\mathbb{R}^m$.

4.1. Parametrized family of acoustic waves. We consider in this section the propagation of a parametrized family of acoustic waves:

$$\tag{4.1}\begin{aligned}
&\partial_t^2 u^\varepsilon - a^\varepsilon(y)\,\Delta_x u^\varepsilon = 0, \quad t\in]0,T[, \quad (x,y)\in\Omega_x\times\mathcal{O},\\
&u^\varepsilon\,|_{\partial\Omega_x\times\mathcal{O}} = 0, \qquad t\in]0,T[,\\
&u^\varepsilon\,|_{t=0} = \alpha, \qquad \partial_t u^\varepsilon\,|_{t=0} = \beta \quad \text{in } \Omega_x\times\mathcal{O}.
\end{aligned}$$

Here (a^ε) is a sequence in $L^\infty(\mathcal{O})$ that satisfies

$$\tag{4.2}\begin{aligned}
&a^\varepsilon(y)\in\overline{\Lambda} \quad \text{a.e. in } \mathcal{O}, \quad \Lambda =]a_-,a_+[, \quad 0 < a_- \le a_+,\\
&a^\varepsilon \overset{\star}{\rightharpoonup} \overline{a} \quad \text{in } L^\infty(\mathcal{O}) \text{ weak*}.
\end{aligned}$$

The initial data are taken such that

$$\tag{4.3} \alpha\in L^2(\mathcal{O}; H_0^1(\Omega_x)), \qquad \beta\in L^2(\mathcal{O}; L^2(\Omega_x)).$$

We state and prove the following result.

THEOREM 4.1. *Suppose* (4.2), (4.3) *hold. Then, along a subsequence, there exists a parametrized family of nonnegative measures* $d\sigma_y$, *with support in* $\overline{\Lambda}$, *such that the sequence* (u^ε) *of solutions of* (4.1) *converges in* $L^\infty(0,T;L^2(\mathcal{O};H_0^1(\Omega_x)))$ *weak* and in* $H^1(0,T;L^2(\mathcal{O};H_0^1(\Omega_x)))$ *weak to* u *solution, in the sense of distributions, of the following system:*

$$\tag{4.4}\begin{aligned}
&\partial_t^2 u - \overline{a}(y)\Delta_x u + \int_\Lambda d\sigma_y(\lambda)\Delta_x w(\lambda) = 0 \quad \text{in }]0,T[\,\times\Omega_x\times\mathcal{O},\\
&\partial_t^2 w - \lambda\Delta_x w + \Delta_x u = 0 \quad \text{in }]0,T[\,\times\Omega\times\Lambda,\\
&u\,|_{\partial\Omega} = 0, \qquad w\,|_{\partial\Omega_x\times\mathcal{O}\times\Lambda} = 0 \quad \text{in }]0,T[,\\
&u\,|_{t=0} = \alpha, \qquad \partial_t u\,|_{t=0} = \beta \quad \text{in } \Omega,\\
&w\,|_{t=0} = 0, \qquad \partial_t w\,|_{t=0} = 0 \quad \text{in } \Omega\times\Lambda.
\end{aligned}$$

Remark 4.2. In the case of the Cauchy problem ($\Omega_x = \mathbb{R}^n$) the function u satisfies in $]0,T[\times\mathbb{R}^n \times \mathcal{O}$ the following nonlocal wave equation:

$$\tag{4.5} \begin{aligned} &\partial_t^2 u - \bar{a}(y)\Delta_x u + \int_0^t \int_{\mathbb{R}^n} ds\, M_n(t-s, x-z, y)\Delta_x^2 u(s,z,y)(x)\, ds\, dz = 0,\\ &u\,|_{t=0} = \alpha, \qquad \partial_t u\,|_{t=0} = \beta. \end{aligned}$$

The kernel $M_n(t,x,y)$ is given by $M_n(t,x,y) = \int_\Lambda d\sigma_y(\lambda)\, \mathcal{E}_n(t,x,\lambda)$, where $\mathcal{E}_n(\cdot,\cdot,\lambda)$ is the elementary solution in $\mathbb{R}^n$ of the wave operator $\partial_t^2 - \lambda\,\Delta_x$, $\lambda \in \Lambda$.

Proof of Theorem 4.1. Let $\mathcal{L}$ denote the Laplace transform in time and $v^\varepsilon(p,x,y) = \mathcal{L}(u^\varepsilon(\cdot,x,y))(p)$. Then, for fixed $p \in \mathbb{C}$ with $\Re e\, p > p_0 > 0$, v^ε is the unique solution of the Dirichlet problem

$$\tag{4.6} p^2\, v^\varepsilon - a^\varepsilon(y)\,\Delta_x v^\varepsilon = F \quad \text{in } \Omega_x \times \mathcal{O}, \quad v^\varepsilon\,|_{\partial\Omega} = 0,$$

where the source term is given by $F(p,x,y) = p\,\alpha(x,y) + \beta(x,y)$. Since the function $F(p,\cdot,\cdot)$ belongs to $L^2(\mathcal{O}; L^2(\Omega_x))$ then, according to the regularity properties of elliptic equations, $v^\varepsilon(p,\cdot,y)$ belongs to $L^2(\mathcal{O}; H_0^1(\Omega_x) \cap H^2(\Omega_x))$. So, let us introduce the unbounded operator A in $L^2(\Omega_x)$ defined by

$$\tag{4.7} A\,u = -\Delta_x u \quad \text{with domain } D(A) = H_0^1(\Omega_x) \cap H^2(\Omega_x).$$

Hence, from (4.6), one deduces that for almost every $y \in \mathcal{O}$,

$$\tag{4.8} v^\varepsilon(p,\cdot,y) = (p^2 + a^\varepsilon(y)A)^{-1} F(p,\cdot,y).$$

Using the Dunford–Taylor integral representation, one has

$$\tag{4.9} v^\varepsilon(p,\cdot,y) = \left(\frac{-1}{2\,\imath\,\pi}\int_{\Gamma_p} (p^2 - a^\varepsilon(y)z)^{-1}\,(z+A)^{-1}\,dz\right)(F(p,\cdot,y)),$$

where

$$\Gamma_p = \left\{ z \in \mathbb{C};\ \frac{p^2}{z} \in \tilde{\Gamma}_p \right\}$$

with $\tilde{\Gamma}_p$ a closed curve in the right half space $\Re e\, z > 0$ containing the real interval $\overline{\Lambda}$. Consider now for fixed z in Γ_p the sequence $(H^\varepsilon(\cdot,p,z))$ defined for almost every $y \in \mathcal{O}$ by $H^\varepsilon(y,p,z) = (p^2 - a^\varepsilon(y)z)^{-1}$. Thanks to Lemma 2.5, there exists a parametrized family of nonnegative measures $d\sigma_y(\cdot)$ supported, for almost every $y \in \mathcal{O}$, in $\overline{\Lambda}$ such that for a subsequence,

$$H^\varepsilon(\cdot,p,z) \overset{\star}{\rightharpoonup} H(\cdot,p,z) \quad \text{in } L^\infty(\mathcal{O}) \text{ weak*},$$

with

$$\tag{4.10} H(y,p,z) = \left(p^2 - z\,\bar{a}(y) - z^2\int_\Lambda (p^2 - \lambda\,z)^{-1}\,d\sigma_y(\lambda)\right)^{-1}.$$

Hence, for a subsequence,

$$v^\varepsilon(p,\cdot,\cdot) \rightharpoonup v(p,\cdot,\cdot) \quad \text{in } L^2(\mathcal{O}; H_0^1(\Omega_x)) \text{ weak},$$

and

$$v(p,\cdot,y) = \frac{-1}{2\imath\pi}\int_{\Gamma_p} H(y,p,z)\,(z+A)^{-1}(F(p,\cdot,y))\,dz \quad \text{a.e. in } \mathcal{O}. \tag{4.11}$$

Plugging (4.10) into (4.11) and using Dunford–Taylor's integral, we find

$$v(p,\cdot,y) = \left(p^2 + \bar{a}(y)\,A - \int_\Lambda A(p^2+\lambda A)^{-1}A\,d\sigma_y(\lambda)\right)^{-1}(F)\,(p,\cdot,y). \tag{4.12}$$

We now introduce the new function $w_*(p,x,y,\lambda)$ defined for $\Re e p > p_0 > 0$, $\lambda \in \Lambda$ and for almost every $y \in \mathcal{O}$ by

$$w_*(p,\cdot,y,\lambda) = (p^2+\lambda A)^{-1}Av(p,\cdot,y). \tag{4.13}$$

Clearly w_* is the unique solution of the Dirichlet problem

$$p^2\,w_* - \lambda\,\Delta_x w_* = -\Delta_x v \quad \text{in } \Omega_x\times\mathcal{O}\times\Lambda, \qquad w_*\,|_{\partial\Omega_x\times\mathcal{O}\times\Lambda} = 0. \tag{4.14}$$

Equations (4.12) and (4.14) imply that (v, w_*) is a solution for $p \in \mathbb{C}$, $\Re e p > p_0$ of the following system:

$$\begin{aligned} &p^2\,v - \bar{a}(y)\Delta_x v + \int_\Lambda d\sigma_y(\lambda)\,\Delta_x w_* = F(p,\cdot,y) \quad \text{in } \Omega,\\ &p^2\,w_* - \lambda\,\Delta_x w_* + \Delta_x v = 0 \quad \text{in } \Omega_x\times\mathcal{O}\times\Lambda,\\ &v\,|_{\partial\Omega_x\times\mathcal{O}} = 0, \qquad w_*\,|_{\partial\Omega_x\times\mathcal{O}\times\Lambda} = 0. \end{aligned} \tag{4.15}$$

Let $w(t,x,y,\lambda)$ be defined in $]0,T[\times\Omega_x\times\mathcal{O}\times\Lambda$ by its Laplace transform w_*:

$$\mathcal{L}(w(\cdot,x,y,\lambda))(p) = w_*(p,x,y,\lambda). \tag{4.16}$$

One deduces from (4.15) and (4.16) that (u,w) is a solution of system (4.4) and proves the theorem.

Remark 4.3. We can study, along lines similar to §§2 and 3, (4.1) with a source term of type

$$\begin{aligned} &f^\varepsilon(t,x,y) = c^\varepsilon(y)\,f(t,x),\\ &\qquad f \in L^2(]0,T[\times\Omega_x), \qquad 0 < c_- \le c^\varepsilon(y) \le c_+ \quad \text{a.e. in } \mathcal{O}. \end{aligned} \tag{4.17}$$

4.2. A damped wave equation. This second part is concerned with the study of some kind of interaction of oscillations by the homogenization process in the damped wave equation for which the damping acts on the transverse variable $y \in \mathcal{O}$. We consider the Dirichlet problem

$$\begin{aligned} &\rho^\varepsilon(x)\,\partial_t^2 u^\varepsilon - \mathrm{div}_x(k^\varepsilon(x)\ \mathrm{grad}_x u^\varepsilon) + \theta^\varepsilon(y)\,\partial_t u^\varepsilon = 0,\\ &u^\varepsilon\,|_{\partial\Omega_x\times\mathcal{O}} = 0, \quad u^\varepsilon\,|_{t=0} = \alpha, \quad \partial_t u^\varepsilon\,|_{t=0} = \beta, \end{aligned} \tag{4.18}$$

where $\alpha \in L^2(\mathcal{O}; H_0^1(\Omega_x))$ and $\beta \in L^2(\mathcal{O}; L^2(\Omega_x))$. We assume that ρ^ε belongs to $L^\infty(\Omega_x)$ and k^ε is a symmetric tensor with components in $L^\infty(\Omega_x)$ satisfying, for almost every $x \in \Omega_x$,

$$0 < \rho_- \le \rho^\varepsilon(x) \le \rho_+, \tag{4.19}$$

$$k_-|\xi|^2 \le k^\varepsilon(x)\xi\cdot\xi \le k_+|\xi|^2 \quad \forall\xi\in\mathbb{R}^n, \tag{4.20}$$

with $0<k_- \le k_+$. Suppose also that

$$\begin{aligned} &\rho^\varepsilon \overset{\star}{\rightharpoonup} \overline{\rho} \quad \text{in } L^\infty(\Omega_x) \text{ weak*},\\ &k^\varepsilon \overset{\mathcal{H}}{\rightharpoonup} \overline{k} \quad \text{in the sense of homogenization [15], [19].}\end{aligned} \tag{4.21}$$

The sequence $(\theta^\varepsilon(y))$ is in $L^\infty(\mathcal{O})$ such that

$$\begin{cases} 0<\theta_- \le \theta^\varepsilon(y) \le \theta_+ \quad \text{a.e. for } y\in\mathcal{O},\\ \theta^\varepsilon \overset{\star}{\rightharpoonup} \overline{\theta} \quad \text{in } L^\infty(\mathcal{O}) \text{ weak*}.\end{cases} \tag{4.22}$$

Let $p\in\mathbb{C}$ with $\Re e\, p>p_0$ for some $p_0>0$. We denote by H_p^ε the unbounded operator in $L^2(\Omega_x)$, with domain $D(H_p^\varepsilon)=\{v\in H_0^1(\Omega_x);\ H_p^\varepsilon v\in L^2(\Omega_x)\}$ defined by

$$H_p^\varepsilon(v)=\rho^\varepsilon(x)\,p^2\,v-\operatorname{div}_x(k^\varepsilon(x)\operatorname{grad}_x v), \qquad v\in D(H_p^\varepsilon). \tag{4.23}$$

Taking the Laplace transform in time of (4.18), the function $v^\varepsilon(p,x,y)=\mathcal{L}(u^\varepsilon(\cdot,x,y))(p)$ is then the unique solution of the equation

$$\left(H_p^\varepsilon+p\,\theta^\varepsilon(y)\right)v^\varepsilon=F^\varepsilon(p,x,y)+\theta^\varepsilon(y)\,\alpha(x,y),$$

where $F^\varepsilon(p,x,y)=\rho^\varepsilon(x)\,(p\,\alpha(x,y)+\beta(x,y))$. Arguing as in the proof of Theorem 3.4, we must pass to the limit in the following two sequences: $(H_p^\varepsilon+p\,\theta^\varepsilon(y))^{-1}(F^\varepsilon)$ and $(H_p^\varepsilon+p\,\theta^\varepsilon(y))^{-1}(\theta^\varepsilon(y)\,\alpha)$. The resolvent operator $(H_p^\varepsilon+p\,\theta^\varepsilon(y))^{-1}$, defined for $\Re e\,p>p_0$, have the following integral representation:

$$\left(H_p^\varepsilon+p\,\theta^\varepsilon(y)\right)^{-1}=\frac{-1}{2\,\imath\,\pi}\int_{\Gamma_p}(z-p\,\theta^\varepsilon(y))^{-1}\,(H_p^\varepsilon+z)^{-1}\,dz.$$

Here

$$\Gamma_p=\left\{z\in\mathbb{C};\frac{z}{p}\in\tilde{\Gamma}_p\right\}$$

with $\tilde{\Gamma}_p$ a closed curve in the right half plane $\Re e\, z>0$ containing the real interval $\overline{\Lambda}$. Thanks to Lemma 2.5, one has, for a subsequence

$$\left(z-p\,\theta^\varepsilon(\cdot)\right)^{-1}\overset{\star}{\rightharpoonup}\left(z-p\overline{\theta}(\cdot)-p^2\int_\Lambda(z-p\lambda)^{-1}\,d\sigma_\bullet(\lambda)\right)^{-1}(F) \tag{4.24}$$

in $L^\infty(\mathcal{O})$ weak*, where $d\sigma_y(\cdot)$ is a parametrized family of nonnegative measures associated with (θ^ε). From (4.21), (F^ε) converges weakly in $L^2(\Omega)$ to F, where $F(p,x,y)=\overline{\rho}(x)\,(p\,\alpha(x,y)+\beta(x,y))$. We claim that, for all $z\in\Gamma_p$,

$$(H_p^\varepsilon+z)^{-1}(F^\varepsilon(p,\cdot,\cdot))\ \to\ (H_p+z)^{-1}(F(p,\cdot,\cdot)) \quad \text{strongly in } L^2(\Omega), \tag{4.25}$$

where H_p is the homogenized operator of H_p^ε given by

$$\begin{aligned} &D(H_p)=\{v\in H_0^1(\Omega_x);\ H_p v\in L^2(\Omega_x)\},\\ &H_p(v)=\overline{\rho}(x)\,p^2\,v-\operatorname{div}_x(\overline{k}(x)\operatorname{grad}_x v) \quad \text{for } v\in D(H_p).\end{aligned} \tag{4.26}$$

For $z \in \Gamma_p$, let χ^ε be the solution of

$$(4.27)\qquad (H_p^\varepsilon + z)\chi^\varepsilon = F^\varepsilon \quad \text{in } \Omega_x \times \mathcal{O}, \quad \chi^\varepsilon|_{\partial\Omega_x \times \mathcal{O}} = 0.$$

The sequence (χ^ε) lies in a bounded set of $L^2(\mathcal{O}; H_0^1(\Omega_x))$. Thus, for a subsequence, (χ^ε) converges weakly in $L^2(\Omega)$ to χ. For a given function φ, defined in Ω, let us denote by $\tilde{\varphi}$ its extension which takes the value zero outside Ω. For h_y given in $\mathbb{R}^m$, we set

$$\delta_{h_y}\tilde{\varphi}(x,y) = \tilde{\varphi}(x, y+h_y) - \tilde{\varphi}(x,y).$$

Clearly, $\delta_{h_y}\tilde{\chi}^\varepsilon(p,\cdot,y)$ is, for almost every $y \in \mathbb{R}^m$, the unique solution of

$$(H_p^\varepsilon + z)\,\delta_{h_y}\tilde{\chi}^\varepsilon(p,\cdot,y) = \rho^\varepsilon(\cdot)\,\delta_{h_y}\tilde{F}(p,\cdot,y) \quad \text{in } \Omega_x,$$

$$\delta_{h_y}\tilde{\chi}^\varepsilon(p,\cdot,y)\,|_{\partial\Omega_x} = 0,$$

and satisfies the standard estimate

$$\|\delta_{h_y}\tilde{\chi}^\varepsilon(p,\cdot,y)\|_{H_0^1(\Omega_x)} \le C\,|\delta_{h_y}\tilde{F}(p,\cdot,y)|_{L^2(\Omega_x)} \quad \text{for almost every } y \in \mathbb{R}^m,$$

where $C > 0$ is independent of ε and h_y. One easily deduces that, for $h_x \in \mathbb{R}^n$, $h_y \in \mathbb{R}^m$, $|h_x|$ small,

$$\int_{\Omega_x \times \mathcal{O}} |\tilde{\chi}^\varepsilon(p, x+h_x, y+h_y) - \tilde{\chi}^\varepsilon(p,x,y)|^2 dx$$
$$\le C\left(|h_x|^2\,|F|^2_{L^2(\Omega_x\times\mathcal{O})} + |\delta_{h_y}\tilde{F}(p,\cdot,\cdot)|^2_{L^2(\Omega_x\times\mathcal{O})}\right),$$

with $C > 0$ independent of ε, h_x, and h_y. Thus the sequence (χ^ε) is compactly imbedded in $L^2(\Omega_x \times \mathcal{O})$. Hence, there exists a subsequence (still denoted by χ^ε) such that $\chi^\varepsilon \to \chi$ strongly in $L^2(\Omega)$. Using the standard arguments of the homogenization theory, see Murat [15] and Tartar [19], one verifies that $\chi(p,\cdot,y)$ is the unique solution in $H_0^1(\Omega_x)$ of the homogenized problem

$$\overline{\rho}(x)\,p^2\,\chi - \mathrm{div}_x(\overline{k}(x)\mathrm{grad}_x\chi) + z\,\chi = F(p,\cdot,y) \quad \text{for almost every } y \in \mathcal{O}.$$

This proves (4.25). Consequently, for all $z \in \Gamma_p$,

$$(z - p\,\theta^\varepsilon(y))^{-1}\,(H_p^\varepsilon + z)^{-1}(F^\varepsilon) \rightharpoonup \left(z - p\overline{\theta} - \int_\Lambda d\sigma_y(\lambda)\,(z - p\,\lambda)^{-1}\right)^{-1}(H_p + z)^{-1}(F)$$

weakly in $L^2(\Omega)$. Let $\psi \in L^2(\Omega)$ fixed and define

$$m^\varepsilon(z) = \int_\Omega \psi(x,y)\,(z - p\,\theta^\varepsilon(y))^{-1}\,(H_p^\varepsilon + z)^{-1}(F^\varepsilon(p,x,y))\,dx\,dy,$$

$$m(z) = \int_\Omega \psi(x,y)\left(z - p\overline{\theta}(y) - \int_\Lambda d\sigma_y(\lambda)\,(z - p\,\lambda)^{-1}\right)^{-1}(H_p + z)^{-1}F(p,x,y)\,dx\,dy.$$

We have proved that $m^\varepsilon(z) \to m(z)$ for all $z \in \Gamma_p$. In order to apply the dominated convergence theorem to the sequence $\left(\int_{\Gamma_p} m^\varepsilon(z)\,dz\right)$, see Folland [11, p. 157] for instance, let us prove that $m^\varepsilon(z)$ is bounded by an integrable function on Γ_p. The strong convergence of the sequence $(H_p^\varepsilon + z)^{-1}(F^\varepsilon)$ in $L^2(\Omega)$, the energy estimate

$$|(H_p^\varepsilon + z)^{-1}(F^\varepsilon)|_{L^2(\Omega)} \leq C\,|F|_{L^2(\Omega)},$$

with C independent of ε and z, and the dominated convergence theorem yield

$$\int_{\Gamma_p} |(H_p^\varepsilon + z)^{-1}(F^\varepsilon)|_{L^2(\Omega)}\,dz \to \int_{\Gamma_p} |(H_p + z)^{-1}(F)|_{L^2(\Omega_x \times \mathcal{O})}\,dz.$$

Since $\tilde{\Gamma}_p$ is a closed curve in $\mathbb{C}$ containing $\overline{\Lambda}$, $\Lambda =]\theta_-, \theta_+[$, there exists $r > 0$ such that

$$(z - p\,\theta^\varepsilon(y)) \geq r > 0 \quad \text{for all } z \in \tilde{\Gamma}_p \quad \text{and for almost every } y \in \mathcal{O}.$$

Using the Cauchy–Schwarz inequality, one gets

$$|m^\varepsilon(z)| \leq \frac{C}{r}|\psi|_{L^2(\Omega)}\,|(H_p^\varepsilon + z)^{-1}(F^\varepsilon)|_{L^2(\Omega)}.$$

From the dominated convergence theorem, it follows that, for all $\psi \in L^2(\Omega)$,

$$\int_{\Gamma_p} m^\varepsilon(z)\,dz \to \int_{\Gamma_p} m(z)\,dz.$$

This convergence means that the sequence $(\int_{\Gamma_p}(z - p\,\theta^\varepsilon(\cdot))^{-1}\,(H_p^\varepsilon + z)^{-1}\,dz\,(F^\varepsilon))$ converges weakly in $L^2(\Omega)$ to

$$\int_{\Gamma_p}\left(z - p\bar{\theta}(\cdot) - \int_\Lambda d\sigma_\bullet(\lambda)\,(z - p\,\lambda)^{-1}\right)^{-1}\,(H_p + z)^{-1}(F)\,dz.$$

We finally use Dunford–Taylor's integral to deduce the following weak convergence in $L^2(\Omega)$:

$$\text{(4.28)}\quad (H_p^\varepsilon + p\,\theta^\varepsilon(\cdot))^{-1}(F^\varepsilon) \rightharpoonup \left(H_p + p\bar{\theta}(\cdot) - p^2\int_\Lambda (H_p + \lambda p)^{-1}\,d\sigma_\bullet(\lambda)\right)^{-1}(F).$$

Consider now the sequence $((H_p^\varepsilon + p\,\theta^\varepsilon(y))^{-1}(\theta^\varepsilon(y)\,\alpha))$. Using (3.36) and Lemma 2.5, one proves that, for a subsequence,

$$\text{(4.29)}\qquad (H_p^\varepsilon + p\,\theta^\varepsilon(\cdot))^{-1}(\theta^\varepsilon(\cdot)\,\alpha) \rightharpoonup \frac{-1}{2\,\imath\,\pi}\int_{\Gamma_p}\frac{1}{p}G(p,z,\cdot)\,(H_p + z)^{-1}(\alpha)\,dz$$

in $L^2(\Omega)$ weak, where

$$G(p,z,y) = -1 + z\left[z - p\bar{\theta}(y) - p^2\int_\Lambda (z - p\,\lambda)^{-1}\,d\sigma_y(\lambda)\right]^{-1}.$$

This can be written also as

$$\frac{1}{p}G(p,z,y) = \left(z - p\bar{\theta}(y) - p^2\int_\Lambda (z - p\lambda)^{-1}d\sigma_y(\lambda)\right)^{-1}\left(\bar{\theta}(y) + p\int_\Lambda (z - p\lambda)^{-1}\,d\sigma_y(\lambda)\right).$$

By using the Dunford–Taylor integral formula, one gets

$$(H^\varepsilon_p + p\theta^\varepsilon(\cdot))^{-1}(\theta^\varepsilon(\cdot)\,\alpha) \rightharpoonup -\frac{1}{p}G(p,-H_p,\cdot)(\alpha) \quad \text{in } L^2(\Omega) \text{ weak},$$

and

$$-\frac{1}{p}\,G(p,-H_p,y) = \left(H_p + p\overline{\theta}(y) - p^2\int_\Lambda (H_p+p\lambda)^{-1}\,d\sigma_y(\lambda)\right)^{-1}(h(p,H_p,y)), \tag{4.30}$$

where $h(p,H_p,y) = \overline{\theta}(y) - p\int_\Lambda (H_p+\lambda p)^{-1}\,d\sigma_y(\lambda)$. From (4.28)–(4.30) we deduce the following result.

THEOREM 4.4. *Let u^ε be the solution of* (4.18), *under assumptions* (4.19)–(4.22). *Then, along a subsequence, there exists a parametrized family of nonnegative measures $d\sigma_y$, with support in $\overline{\Lambda}$, such that the sequence (u^ε) converges, as $\varepsilon \to 0$, weakly* in the space $L^\infty(0,T;L^2(\mathcal{O};H^1_0(\Omega_x)))$ and in $H^1(0,T;L^2(\mathcal{O};H^1_0(\Omega_x)))$ weak to u solution, in the sense of distributions, of the following system:*

$$\begin{aligned}
&\overline{\rho}(x)\,\partial_t^2 u - \operatorname{div}_x(\overline{k}(x)\operatorname{grad}_x u) + \overline{\theta}(y)\partial_t u - \int_\Lambda d\sigma_y(\lambda)\partial_t w(\lambda) = 0,\\
&\overline{\rho}(x)\partial_t^2 w - \operatorname{div}_x(\overline{k}(x)\operatorname{grad}_x w) + \lambda\partial_t w - \partial_t u = 0,\\
&u\,|_{\partial\Omega_x\times\mathcal{O}} = 0, \qquad w\,|_{\partial\Omega_x\times\mathcal{O}\times\Lambda} = 0,\\
&u\,|_{t=0} = \alpha, \quad w\,|_{t=0} = 0, \quad \partial_t u\,|_{t=0} = \beta, \quad \partial_t w\,|_{t=0} = 0,
\end{aligned} \tag{4.31}$$

where $t \in]0,T[$, $x \in \Omega_x$, $y \in \mathcal{O}$, $\lambda \in \Lambda$.

REFERENCES

[1] N. I. AHIEZER AND M. G. KREIN, *Some Questions in the Theory of Moments*, Translations of Math. Monographs 2, American Mathematical Society, Providence, RI, 1962.

[2] G. ALLAIRE, *Homogénéisation et convergence à deux échelles. Application à un problème de diffusion turbulente*, Note C. R. Acad. Sci., Paris, Sér. I Math., 312 (1991), pp. 581–586.

[3] Y. AMIRAT, K. HAMDACHE, AND A. ZIANI, *Homogénéisation d'équations hyperboliques du premier ordre et application aux écoulements miscibles en milieu poreux*, Ann. Inst. Henri Poincaré, Anal. non linéaire, 6 (1989), pp. 397–417.

[4] ———, *Some results on homogenization of convection-diffusion equations*, Arch. Rational Mech. Anal., 114 (1991), pp. 155–178.

[5] ———, *Homogenization of parametrized families of hyperbolic problems*, Proc. Roy. Soc. Edinburgh, 120 A (1992), pp. 199–221.

[6] ———, *Homogénéisation non locale pour des équations dégénérées à coefficients périodiques*, Note C. R. Acad. Sc., Paris, Sér. I Math., 312 (1991), pp. 963–966.

[7] ———, *Singular Perturbations and Periodic Homogenization*, Rapport 136, Equipe d'Analyse Numérique Lyon Saint-Etienne, 1992.

[8] A. BENSOUSSAN, J. L. LIONS, AND G. C. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.

[9] W. F. DONOGHUE, *Monotone Matrix Functions and Analytic Continuation*, Springer-Verlag, New York, 1974.

[10] W. E, *Homogenization of Linear and Nonlinear Transport Equations*, Comm. Pure Appl. Math., 45 (1992), pp. 301–326.

[11] G. FOLLAND, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1976.

[12] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[13] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et Applications, Vol.* I, Dunod, Paris, 1968.

[14] S. Mizohata, *The Theory of Partial Differential Equations*, Cambridge University Press, London, 1973.

[15] F. Murat, *H-convergence, Séminaire Faculté des Sciences,* Univeristy of Algiers, Algeria, 1978.

[16] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.

[17] ———, *Asymptotic analysis for a stiff variational problem arising in Mechanics*, SIAM J. Math. Anal., 21 (1990), pp. 1394–1414.

[18] E. Sanchez-Palencia, *Non-Homogeneous Media and Vibration Theory*, Lectures Notes in Phys. 127, Springer-Verlag, New York, 1980.

[19] L. Tartar, *Problèmes d'homogénéisation dans les équations aux dérivées partielles,* Cours Peccot, Collège de France, 1977 (partially written in [15]).

[20] ———, *Compensated Compactness and Applications to* P.D.E., in Research Notes in Math., Non Linear Analysis and Mechanics, Heriot-Watt Symposium, 39, R. J. Knops, ed., Pitman Press, Boston, MA, 1979.

[21] V. N. Vragov, *On a mixed problem for a class of hyperbolic-parabolic equations*, Soviet Math. Dokl., 16 (1975), pp. 1179–1183.

SIAM J. MATH. ANAL.
Vol. 24, No. 5, pp. 1254–1276, September 1993

REGIONAL BLOW UP IN A SEMILINEAR HEAT EQUATION WITH CONVERGENCE TO A HAMILTON–JACOBI EQUATION*

VICTOR A. GALAKTIONOV[†] AND JUAN L. VAZQUEZ[‡]

Abstract. The authors investigate the asymptotic behaviour of blowing-up solutions $u = u(x,t) \geqq 0$ to the semilinear parabolic equation with source

$$u_t = u_{xx} + (1+u)\log^2(1+u) \quad \text{for } x \in \mathbf{R},\ t > 0,$$

with nonnegative and radial symmetric initial data $u_0(|x|)$ that are nonincreasing in $|x|$. Any nontrivial solution u to this problem blows up in a finite time $T > 0$. It is remarkable that the blow-up behaviour of u as t approaches T can be described by the exact blow-up solutions of the quasilinear Hamilton–Jacobi equation

$$U_t = \frac{(U_x)^2}{1+U} + (1+U)\log^2(1+U),$$

with the same blow-up time T. These explicit profiles are only approximate solutions for the problem. The authors prove that the blow-up set B of the solution satisfies meas $(B) \geqq 2\pi$, and under some additional hypothesis on the initial function it is shown that B is just the interval $[-\pi, \pi]$ and the rescaled blow-up shape consists of one hump with formula $\cos^2(x/2)$. The proofs rely on the knowledge of a family of explicit solutions, the method of intersection comparison, some dynamical systems ideas, and a stability analysis for solutions of the Hamilton–Jacobi equation.

Key words. semilinear heat equation, regional blow up, asymptotic behaviour, Hamilton–Jacobi equation

AMS subject classifications. 35K55, 35K65, 34E10

1. Introduction.

In this paper we consider the Cauchy problem for the semilinear heat equation

$$u_t = u_{xx} + (\ 1+u)\log^2(1+u) \quad \text{for } x \in \mathbf{R},\ t > 0, \tag{1.1}$$

$$u(x,0) = u_0(x) \quad \text{for } x \in \mathbf{R}. \tag{1.2}$$

We assume that the initial function $u_0 \in L^1_{\text{loc}}(\mathbf{R}) \cap L^\infty(\mathbf{R})$ is radially symmetric, $u_0 = u_0(|x|)$, nonnegative, $u_0 \geqq 0$, and nonconstant. Our main results also assume that u_0 is nonincreasing as a function of $|x| > 0$. The final section deals with periodic solutions.

Problem (1.1)–(1.2) has a unique classical solution $u = u(|x|,t) > 0$ defined in some maximal time interval [Fr1]. Since the heat source $Q(u) = (1+u)\log^2(1+u)$ in the right-hand side of (1.1) satisfies $Q(u) > 0$ for $u > 0$, $Q(u) = u^2 + o(u^2)$ as $u \to 0$ and $\int_1^\infty dz/Q(z) < \infty$, the solution blows up in a finite time T, $0 < T < \infty$ (cf. [Fu]): $u(x,t)$ is a classical solution of (1.1) in $\mathbf{R} \times (0,T)$ and

$$\sup_{x \in \mathbf{R}} u(x,t) \to \infty \quad \text{as } t \to T. \tag{1.3}$$

Much work has been devoted in recent years to understanding finite-time blow up for equations of the form $u_t = u_{xx} + Q(u)$. Four major problems arise in this context and have been studied by different authors; see the books [BE] and [SGKM] for extensive

* Received by the editors April 20,1992; accepted for publication (in revised form) December 3, 1992. This work was partially supported by EEC contract SC1-0019-C(TT).

[†] Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Miusskaya Square 4, 125047, Moscow, Russia.

[‡] Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain.

references. They are: (i) existence of blow up in terms of the initial data, (ii) the form of the *blow-up set*

$$B = B(u_0) \equiv \{x \in \mathbf{R} : \exists x_n \to x \text{ and } t_n \to T \text{ such that } u(x_n, t_n) \to \infty \text{ as } n \to \infty\}, \tag{1.4}$$

(iii) the rate of divergence of $u(x,t)$ as $t \to T$ for points $x \in B$, and (iv) the shape of the solution at $t = T$, after scaling out this rate. The aim of the paper is to answer these questions for the solutions of (1.1) for different choices of initial data.

The simplest form of blow up is *global flat blow up*, i.e., the solution blows up in the whole space with no spatial structure. This blow-up form can always be obtained by starting with a constant initial function, and the precise profile $u = u(t)$ is obtained by integrating the ordinary differential equation (ODE): $u_t = Q(u)$. It is remarkable that precisely the opposite situation occurs for many blow-up equations for suitably concentrated initial data, say when u_0 is bell-shaped with one maximum and compact support: then blow up occurs at an isolated point, *single-point blow up*. Other solutions many blow up at a finite number of points. This happens, for instance, in the most studied cases $Q(u) = u^p$ with $p > 1$ and $Q(u) = \exp(u)$. See, for instance, [FM], [CM], [V].

Since no blow up occurs for $Q(u) = u^p$ with $p = 1$, it is interesting to understand the limit situation $p \approx 1$. In this direction the family of equations

$$u_t = u_{xx} + (1+u)\log^{\beta}(1+u), \qquad \beta > 0 \tag{1.5}$$

was introduced in [GKMS] in 1979 (see also [Sam], generalizations in [G1], and the references of [SGKM, Chap. IV]). The case $\beta = 2$ has a particular interest since it serves as a limit case for the blow-up behaviour as we explain below.

Indeed, qualitative and numerical methods described in the above references had suggested that for suitably concentrated initial data blow up occurs in problem (1.1), (1.2) in a bounded domain, i.e., we have the so-called *regional blow up*. Moreover, these studies pointed out that the behaviour of the solution to (1.1), (1.2) as it approaches blow up should be described by some solution of the quasilinear Hamilton–Jacobi equation

$$U_t = \frac{(U_x)^2}{1+U} + (1+U)\log^2(1+U) \quad \text{for } x \in \mathbf{R},\ t \in (0,T), \tag{1.6}$$

of the form

$$U(x,t) = \exp\left\{(T-t)^{-1}\,\overline{g}(x)\right\} - 1. \tag{1.7}$$

By substituting (1.7) into (1.6) one can obtain that the function $\overline{g} \geqq 0$ solves the first-order quadratic ordinary differential equation

$$(\overline{g}_x)^2 + \overline{g}^2 - \overline{g} = 0 \quad \text{in } \mathbf{R}. \tag{1.8}$$

In this paper we give a rigorous proof of these facts for the class of solutions that are radially symmetric in the space variable x and nonincreasing in $|x| > 0$. Let us briefly describe the main ideas and results: in studying the blow-up behaviour of (1.1) it is convenient to introduce the transformation

$$u(x,t) = e^{v(x,t)} - 1, \tag{1.9}$$

which yields the following equation for the function $v(x,t)$:

$$(1.10) \qquad v_t = v_{xx} + (v_x)^2 + v^2 \quad \text{for } x \in \mathbf{R},\ t > 0.$$

In view of the expected behaviour (1.7), we also introduce the rescaled function

$$(1.11) \qquad \theta(x,\tau) = (T-t)v(x,t),$$

where $\tau = -\log(T-t)$ (so that, in particular, the new time $\tau \to \infty$ as $t \to T$). The function $\theta(x,\tau)$ solves the Cauchy problem

$$(1.12) \qquad \theta_\tau = e^{-\tau}\theta_{xx} + (\theta_x)^2 + \theta^2 - \theta \quad \text{for } x \in \mathbf{R},\ \tau > \tau_0 = -\log T,$$

$$(1.13) \qquad \theta(x,\tau_0) = \theta_0(x) \equiv T\log(1+u_0(x)) \quad \text{for } x \in \mathbf{R}.$$

After describing in §2 a family of 2π-periodic blow-up solutions constructed in [GP3], which satisfy (1.10) but blow up in the whole space and are essential in later comparison arguments, we proceed in §§3–8 to study solutions with radially symmetric and nonincreasing initial data. Thus, in §3 we establish a sharp estimate for the supremum of θ as $\tau \to \infty$: $\sup\theta(\cdot,\tau) \to 1$ (Proposition 3.1 and 3.2). Section 4 is devoted to proving a semiconvexity result, of the form $\theta_{xx} \geqq -C$ for τ large (Proposition 4.1), which together with the previous estimate implies that single point blow up cannot occur.

Regional blow up means that $u(x,t) \to \infty$ as $t \to T$ in a set of nonzero finite measure. We have already introduced the blow-up set, where u diverges. A more precise divergence is obtained on a possible smaller set,

$$(1.14) \qquad \begin{aligned} B_* = B_*(u_0) \equiv \{x \in \mathbf{R} : \exists t_n \to T \text{ such that} \\ u(x,t_n) \to \infty \text{ as } n \to \infty\}. \end{aligned}$$

Obviously, $B_* \subseteq B$. By the monotonicity results of [GP1] and [GP4], B_* can also be defined as follows:

$$(1.15) \qquad B_* \equiv \{x \in \mathbf{R} : u(x,t) \to \infty \text{ as } t \to T\}.$$

B_* is called the *monotone blow-up set*; see [G3]. We establish in §5 that the minimal configuration for blow up corresponds to the particular profile

$$(1.16) \qquad G(x) = \begin{cases} \cos^2\left(\dfrac{x}{2}\right) & \text{if } |x| \leqq \pi, \\ 0 & \text{otherwise.} \end{cases}$$

This implies that the blow-up set consists of an interval of length at least 2π. Moreover, $B_* \supseteq (-\pi,\pi)$. Conversely, we obtain in §6 localization of the blow-up set in terms of the number of intersections of u_0 with respect to the explicit solutions of §2 (Theorem 6.1).

In order to better understand the blow-up phenomenon we have to study the stabilization of the function $\theta(x,\tau)$ as $\tau \to \infty$ to a solution $\bar{g}(x)$ of the stationary equation (1.8). This is done in §§7 and 8 where we prove that under a strict condition (which is satisfied, for instance, by suitably bell-shaped initial functions with support in $(-\pi,\pi)$; see condition (7.26)) the blow-up set is exactly $[-\pi,\pi]$ and θ converges as $\tau \to \infty$ to the function G. In these proofs we consider (1.12) as a perturbation of the Hamilton–Jacobi equations

$$(1.17) \qquad g_\tau = (g_x)^2 + g^2 - g \quad \text{in } \mathbf{R} \times (0,\infty),$$

and apply the asymptotic results for small perturbations of autonomous dynamical systems of [GV] and some stability analysis of the viscosity solutions of (1.17). In fact, we will need an improvement of the result of [GV] which simplifies its application.

Many solutions have nonminimal blow up. Solutions with more than one hump can have a nonconnected blow-up set. As an example we consider in §9 x-periodic solutions with period $2m\pi$, m as integer, since we can use the techniques developed in the radially symmetric case. We construct solutions whose blow up consists of the union of intervals of the form $[(2km-1)\pi,\ (2km+1)\pi]$, $k \in \mathbf{Z}$, with a corresponding union of copies of $G(x)$ as an asymptotic profile. We have refrained in this paper from using more general situations for the sake of brevity. We hope to treat some of them, in particular examples of two-hump blow up, in a forthcoming article.

As a final comment on related equations, let us mention that the exponent $\beta = 2$ of our equation is critical for the blow-up behaviour of (1.5). Thus, the Cauchy–Dirichlet Problem has been studied in [GP2] for $\beta > 2$, and *single-point blow up* is shown to occur using the method of [FM], while [L] shows *global blow up* for $1 < \beta < 2$ for the same problem. This leaves $\beta = 2$ as the only case with *regional blow up*. On the other hand, it is interesting to remark that in the case $\beta \in (0,1)$ studied in [GKS1], though solutions are global in time, the asymptotic profiles as $t \to \infty$ are given as the solutions of the Hamilton–Jacobi equations, $U_t = (U_x)^2/(1+U) + (1+U)\log^{\beta}(1+U)$, which parallels (1.6).

Let us also recall that regional blow up has been observed for several models of quasilinear heat equation, like

$$u_t = (u^m)_{xx} + u^p \tag{1.18}$$

for $m > 1$ and $p = m$; see [SGKM, Chap. IV] and [G3]. The present model seems to be the only known case of a *semilinear* heat equation with regional blow up.

2. Explicit solutions. Equation (1.10) admits an explicit solution, 2π-periodic in x, of the form (cf. [GP3], [G2])

$$v_*(x,t) = \varphi(t)[\psi(t) + \cos(x)], \tag{2.1}$$

where the functions $\varphi(t)$ and $\psi(t)$ satisfy the system of ordinary differential equations

$$\varphi' = -\varphi + 2\varphi^2\psi, \quad \psi' = \varphi + \psi - \varphi\psi^2 \quad \text{for } t > 0. \tag{2.2}$$

There exists a one-parameter family of nonnegative blowing-up solutions (2.1) having a given fixed blow-up time $T \in (0,\infty)$. To see this we observe that system (2.2) is equivalent to the following first-order differential equation

$$\frac{d\psi}{d\varphi} = \frac{\varphi + \psi - \varphi\psi^2}{2\varphi^2\psi - \varphi}, \quad \varphi > 0,\ \psi \geqq 1. \tag{2.3}$$

Equation (2.3) admits a one-parameter family of trajectories with the asymptotic behaviour (see [G2])

$$\psi = 1 + \frac{\log\varphi}{2\varphi} + \frac{\mu}{\varphi} + o\left(\frac{1}{\varphi}\right) \quad \text{as } \varphi \to \infty, \tag{2.4}$$

where $\mu \in \mathbf{R}$ is a parameter describing this family. Given a certain μ, namely, on a given trajectory, fixing the blow-up time $T > 0$ is equivalent to fixing an initial point (φ_0, ψ_0)

on the trajectory, corresponding to $t = 0$. It follows from the detailed study of the system (2.2) that for fixed $T > 0$ there exists a constant μ_T such that for any $\mu \geqq \mu_T$ the explicit solution $v_* = v_*(x, t;\, T, \mu)$ with blow-up time T is well defined and nonnegative in $\mathbf{R} \times (0, T)$, while for $\mu < \mu_T$ the initial value of v_* has changing sign. In terms of the functions $\varphi(t), \psi(t)$ the constant μ_T is such that $\psi_0 = 1$ and $\varphi_0 > 0$. This implies that the corresponding initial function has the form

$$v_*(x, 0;\, T, \mu_T) \equiv \varphi_0[1 + \cos(x)] \geqq 0 \quad \text{in } \mathbf{R}. \tag{2.5}$$

Clearly, for $\mu > \mu_T$ we have $\psi_0 > 1$; hence

$$v_*(x, 0) \equiv \varphi_0[\psi_0 + \cos x] \tag{2.6}$$

is positive in $\mathbf{R}$. In any case $v_*(x, t;\, T, \mu) > 0$ in $\mathbf{R} \times (0, T)$ for $\mu \geqq \mu_T$. Integrating the first equation of (2.2) with the help of expansion (2.4) for $\mu \geqq \mu_T$ we obtain the following asymptotic behaviour of the explicit solutions near a fixed blow-up time $t = T$:

$$\begin{aligned} \varphi(t) &= \tfrac{1}{2}(T-t)^{-1}[1 - \tfrac{1}{2}(T-t)|\log(T-t)|(1+o(1))], \\ \psi(t) &= 1 + (T-t)|\log(T-t)|(1+o(1)) \quad \text{as } t \to T. \end{aligned}$$

Therefore, as $t \to T$,

$$\begin{aligned} v_*(x, t;\, T, \mu) = (T-t)^{-1}\Bigg[&\cos^2\left(\frac{x}{2}\right) + \frac{1}{2}\sin^2\left(\frac{x}{2}\right)(T-t)|\log(T-t)| \\ &+ \left(\frac{1}{4} - \frac{1 + 2\log 2 - 4\mu}{4}\sin^2\left(\frac{x}{2}\right)\right)(T-t)(1+o(1))\Bigg]. \end{aligned} \tag{2.7}$$

At the maxima $x = 2\pi k, k = 0, \pm 1, \ldots$, where $\cos^2(x/2) = 1$, we have as $t \to T$

$$\begin{aligned} \sup_{x \in \mathbf{R}} v_*(x, t;\, T, \mu) &\equiv v_*(2\pi k, t;\, T, \mu) \\ &= (T-t)^{-1}[1 + \tfrac{1}{4}(T-t)(1+o(1))]. \end{aligned} \tag{2.8}$$

At the minima $x = \pi + 2\pi k, k = 0, \pm 1, \ldots$, where $\cos^2(x/2) = 0$, the behaviour as $t \to T$ is quite different:

$$\inf_{x \in \mathbf{R}} v_*(x, t;\, T, \mu) \equiv v_*(\pi(2k+1), t;\, T, \mu)\tfrac{1}{2}|\log(T-t)|(1+o(1)). \tag{2.9}$$

For the rescaled explicit solution

$$\theta_*(x, \tau;\, T, \mu) \equiv (T-t)v_*(x, t;\, T, \mu), \quad \tau = -\log(T-t), \tag{2.10}$$

we have as $\tau \to \infty$

$$\theta_*(x, \tau;\, T, \mu) \equiv \cos^2\left(\frac{x}{2}\right) + \frac{1}{2}\sin^2\left(\frac{x}{2}\right)\tau e^{-\tau} + O(e^{-\tau}), \tag{2.11}$$

$$\theta_*(0, \tau;\, T, \mu) \equiv 1 + \tfrac{1}{4}e^{-\tau}(1+o(1)) \to 1, \tag{2.12}$$

$$\theta_*(\pm\pi, \tau;\, T, \mu) \equiv \tfrac{1}{2}\tau e^{-\tau}(1+o(1)) \to 0. \tag{2.13}$$

Notice that $\theta_*(x, \tau;\, T, \mu)$ converges to $G(x)$ as $\tau \to \infty$ for $|x| \leqq \pi$; cf. [G2].

3. Lower and upper bounds. As a first step in controlling the blow-up behaviour of the solution, we derive in this section upper and lower bounds for the supremum in x of the rescaled solution θ defined by (1.11). We begin with the lower bound.

PROPOSITION 3.1. *For every $\tau > \tau_0$ there holds*

$$\sup_{x\in\mathbf{R}} \theta(x,\tau) \equiv \theta(0,\tau) > 1. \tag{3.1}$$

Proof. Equation (1.10) admits a blowing-up solution of the form

$$\overline{v}(t) = (T-t)^{-1}, \tag{3.2}$$

with the same blow-up time, T, as u. Then for any $t \in [0,T)$, $\overline{v}(t)$ intersects in x the solution $v(x,t)$. Indeed, if for some $t = t_0 \in [0,T)$ we have $\overline{v}(t_0) \geqq v(x,t_0)$ in $\mathbf{R}$, then by using the Strong Maximum Principle [Fr1] we arrive at the conclusion that $\overline{v}$ must blow up before v does. This is in contradiction with our assumption. Hence,

$$\sup_{x\in\mathbf{R}} v(x,t) > \overline{v}(t) \quad \text{for } t \in [0,T), \tag{3.3}$$

thus completing the proof. □

Notice that we do not use any special assumptions on u_0 in proving the result. The proof is based on intersection comparison with a solution which is constant as a function of x, and so (3.3) remains valid in much more general situations, e.g., for any arbitrary nonnegative initial function satisfying $u_0 \to 0$ as $|x| \to \infty$. Then $u(x,t) \to 0$ as $|x| \to \infty$ for any fixed $t \in (0,T)$, and (3.3) yields the lower estimate $\sup_{x\in\mathbf{R}} \theta(x,\tau) > 1$ for $\tau \in [\tau_0, \infty]$.

We proceed now with the upper bound. This estimate is much more difficult to obtain.

PROPOSITION 3.2. *Let u_0 be radially symmetric and nonincreasing as a function of $|x|$. A $\tau \to \infty$ there holds*

$$\sup_{x\in\mathbf{R}} \theta(x,\tau) \equiv \theta(0,\tau) < 1 + 2\tau e^{-\tau}. \tag{3.4}$$

Proof. It consists of several steps. We shall use some ideas of the method of stationary states (cf. [SGKM, Chap. 7] and also [GKS2]).

Step 1. *Stationary solutions.* We begin with some simple properties of the stationary solutions of (1.1). Thus, we look at solutions $U(x;\ \lambda)$ of

$$U_{xx} + (1+U)\log^2(1+U) = 0 \quad \text{for } x > 0 \quad \text{and} \quad x < 0, \tag{3.5}$$

$$U_x(0;\ \lambda) = 0, \qquad U(0;\ \lambda) = \lambda, \tag{3.6}$$

where $\lambda > 0$ is an arbitrary constant. We easily show that a positive solution of (3.5), (3.6) exists in a finite interval $\{|x| < x_0(\lambda)\}$, where the endpoint $x_0(\lambda) > 0$ is the point x at which $U(\pm x;\ \lambda)$ reaches the zero value. We let $U(x;\ \lambda) \equiv 0$ for $|x| \geqq x_0(\lambda)$. Since $U(x,\lambda)$ is monotone for $x > 0$ and $x < 0$, we have the inequality

$$U_{xx} \geqq -(1+U)\log^2(1+U) \geqq -(1+\lambda)\log^2(1+\lambda). \tag{3.7}$$

By integrating twice we obtain the lower bound

$$U(x;\ \lambda) \geqq U_-(x;\ \lambda) \equiv (\lambda - \tfrac{1}{2}(1+\lambda)\log^2(1+\lambda)x^2)_+, \tag{3.8}$$

and, in particular,

$$x_0(\lambda) \geqq x_*(\lambda) \equiv \left(\frac{2\lambda}{(1+\lambda)\log^2(1+\lambda)}\right)^{1/2}. \tag{3.9}$$

Multiplying (3.5) by U_x and integrating over $(0,x), 0 \leqq x < x_0(\lambda)$, yields the identity

$$\begin{aligned}(U_x)^2 = {} & \tfrac{1}{2}(1+\lambda)^2[\log^2(1+\lambda) - \log(1+\lambda) + \tfrac{1}{2}] \\ & -\tfrac{1}{2}(1+U)^2[\log^2(1+U) - \log(1+U) + \tfrac{1}{2}].\end{aligned} \tag{3.10}$$

By integrating (3.10) we obtain that

$$x_0(\lambda) \to 0 \quad \text{as } \lambda \to \infty. \tag{3.11}$$

It also follows from (3.10) that for every fixed $m > 0$ and $x' = x'(\lambda, m) > 0$ on the level set $\{x \in \mathbf{R} : U(x;\ \lambda) = m\}$; then

$$U_x(x';\ \lambda) \to \infty \quad \text{as } \lambda \to \infty. \tag{3.12}$$

Step 2. *Intersection comparison.* The result given below is true for a wide class of quasilinear heat equations with source; see [SGKM, Chaps. 6, 7] and [GKS2].

LEMMA 3.3. *There exists $\lambda_* > 0$ large enough such that for any $\lambda > \lambda_*$,*

$$u(x,t) \geqq U(x;\ \lambda) \quad \textit{in } \mathbf{R} \quad \textit{if } u(0,t) \geqq \lambda. \tag{3.13}$$

Proof. For a fixed $t \in [0,T)$ denote by $N(t;\ \lambda)$ the number of intersections in $\{|x| \leqq x_0(\lambda)\}$ of solutions $u(x,t)$ and $U(x;\ \lambda)$, or, in other words, the number of sign changes in $\{|x| \leqq x_0(\lambda)\}$ of the difference $w(x,t;\ \lambda) \equiv u(x,t) - U(x;\ \lambda)$.

By a standard smoothness result for semilinear heat equations we may conclude that $u(x,t)$ is smooth enough for arbitrarily small $t > 0$. It then follows from (3.11) and (3.12) that for every $\lambda > \lambda_*$ large enough there exist exactly two intersections for $t = t_1$ small enough, i. e.,

$$N(t_1;\ \lambda) = 2 \quad \text{for } \lambda > \lambda_*. \tag{3.14}$$

Since $U(\pm x_0(\lambda);\ \lambda) \equiv 0$ and $u(x,t) > 0$ in $\mathbf{R} \times (0,T)$, by a well-known intersection property we conclude that $N(t;\ \lambda)$ does not increase with time (cf. [A], [GP1], [M], [Sat]). Hence by (3.14) we deduce that

$$N(t;\ \lambda) \leqq 2 \quad \text{for } t \in (t_1, T), \lambda > \lambda_*. \tag{3.15}$$

We now prove that if $t_\lambda \in (t_1, T)$ is such that $u(0, t_\lambda) = \lambda$, then

$$N(t_\lambda;\ \lambda) = 0. \tag{3.16}$$

If on the contrary, $N(t_\lambda;\ \lambda) > 0$ (more exactly, $N(t_\lambda;\ \lambda) \geqq 2$ by symmetry), then one can see that for any small $\delta > 0$ the inequality $N(t_\lambda;\ \lambda - \delta) \geqq 4$ holds, contradicting (3.15).

Thus (3.16) implies that $u(x, t_\lambda) \geqq U(x;\ \lambda)$ in $\mathbf{R}$, and hence (3.13) holds by comparison. □

As a consequence of Lemma 3.3 we obtain a control of the L^1-norm near the blow-up time.

LEMMA 3.4. *As $t \to T$ the integral $\|u(\cdot, t)\|_{L^1(\mathbf{R})}$ diverges. More precisely, we have*

$$\tag{3.17} \begin{aligned} &\|u(\cdot, t)\|_{L^1(\mathbf{R})} \geqq \|U(\cdot; u(0,t))\|_{L^1(\mathbf{R})} \geqq \|U_-(\cdot; u(0,t))\|_{L^1(\mathbf{R})} \\ &\quad = \frac{2}{3}\lambda x_*(\lambda)|_{\lambda=u(0,t)} = \frac{2^{2/3}}{3}\frac{u(0,t)}{\log u(0,t)}(1+o(1)) \to \infty. \end{aligned}$$

Though relation (3.17) is not exact in view of the precise asymptotic results that follow, it gives a correct idea of the relative growth of the L^1 and sup-norms.

Step 3. *Ordinary differential inequality for an energy function.* We now introduce the local weighted energy

$$\tag{3.18} E(t) = \int_{-\pi/2}^{\pi/2} \phi(x)u(x,t)\,dx,$$

where

$$\tag{3.19} \phi(x) = \frac{1}{2}\cos(x) > 0 \quad \text{in } \left(-\frac{\pi}{2}, \frac{\pi}{2}\right),$$

so that $\phi'' \equiv -\phi$ and

$$\tag{3.20} \int_{-\pi/2}^{\pi/2} \phi(x)\,dx = 1.$$

An energy estimate is obtained as follows. We multiply (1.1) by $\phi(x)$ and integrate over $(-\pi/2, \pi/2)$ to obtain

$$\tag{3.21} \frac{dE}{dt} \geqq -E + \int_{-\pi/2}^{\pi/2} \phi(x)(1+u)\log^2(1+u)\,dx,$$

valid for $t \in (0,T)$. Now using Jensen's inequality for the convex function $(1+u)\log^2(1+u)$ and (3.20) we have

$$\tag{3.22} \frac{dE}{dt} \geqq -E + (1+E)\log^2(1+E).$$

(Inequalities of the type (3.22) have been used by many authors for proving global (in time) nonexistence of solutions; see, e.g., the first such result in [K] and references in [SGKM].)

By integrating (3.22) over (t,T) and using the fact that $E(T) = \infty$ (see (3.17)) we arrive at the estimate

$$\tag{3.23} T - t \leqq \int_{E(t)}^{\infty} \frac{dz}{(1+z)\log^2(1+z) - z} \quad \text{as } t \to T.$$

Now using the fact that

$$\int_p^{\infty} \frac{dz}{(1+z)\log^2(1+z) - z} = \frac{1}{\log p}\left[1 + \frac{1}{3}\frac{1}{\log^2 p}(1+o(1))\right]$$

as $p \to \infty$, we obtain the following estimate for E.

LEMMA 3.5. *As $t \to T$ we have*

$$T - t \leqq \frac{1}{\log E(t)} \left[1 + \frac{1}{3} \frac{1}{\log^2 E(t)} (1 + o(1)) \right]. \tag{3.24}$$

Step 4. *End of Proof of Proposition* 3.2. By using (3.13), (3.8), and (3.9) we conclude that as $t \to T$,

$$\begin{aligned} E(t) &\geqq \int_{-\pi/2}^{\pi/2} \phi(x) U(x;\, u(0,t))\, dx \geqq \int_{-x_*(\lambda)}^{x_*(\lambda)} \phi(x) U_-(x;\, u(0,t))\, dx \\ &\geqq \frac{1}{2} \phi(0) \| U_-(\cdot;\, u(0,t)) \|_{L^1(\mathbf{R})}. \end{aligned}$$

It then follows from (3.17) that as $t \to T$,

$$E(t) \geqq \frac{1}{4} \frac{2^{2/3}}{3} \frac{u(0,t)}{\log u(0,t)} (1 + o(1)). \tag{3.25}$$

Together (3.24) and (3.25) yield the following inequality

$$T - t \leqq \frac{1}{\log u(0,t)} \left[1 + \frac{\log \log u(0,t)}{\log u(0,t)} (1 + o(1)) \right] \tag{3.26}$$

as $u(0,t) \to \infty$. This inequality can be transformed into

$$\log u(0,t) \leqq (T-t)^{-1} [1 + (T-t) |\log(T-t)| (1 + o(1))] \tag{3.27}$$

as $t \to T$. The last estimate yields (3.4). □

4. Semiconvexity. In the next five sections u_0 is radially symmetric and nonincreasing in $|x|$. We may also assume that u_0 is not constant. The analysis of the asymptotic behaviour is best done in terms of the variable θ introduced in (1.11). As is usual in Dynamical Systems we introduce the ω-limit set of the solution $\theta(x, \tau)$ as follows:

$$\begin{aligned} \omega(\theta_0) = \{ f \in C(\mathbf{R}) : \exists_{\tau_j \to \infty} \text{ such that } \theta(\cdot, \tau_j) \to f(\cdot) \\ \text{as } j \to \infty \text{ uniformly in any compact subset of } \mathbf{R} \}. \end{aligned} \tag{4.1}$$

A key estimate in controlling the ω-limit and in preventing single-point blow up is the following.

PROPOSITION 4.1. *For any $\varepsilon > 0$ there exists a constant $a_\varepsilon > 0$ such that*

$$\theta_{xx}(x, \tau) \geqq - \left(\frac{1}{2} + \varepsilon \right) - \frac{a_\varepsilon}{(\tau - \tau_1)} \tag{4.2}$$

in $\mathbf{R} \times (\tau_1, \infty)$, where $\tau_1 = \tau_0 + 1$.

Proof. The function $z = \theta_{xx}$ solves a semilinear parabolic equation in $\mathbf{R} \times (\tau_0, \infty)$:

$$z_\tau = e^{-\tau} z_{xx} + 2\theta_x z_x + 2z^2 + (2\theta - 1) z + 2(\theta_x)^2. \tag{4.3}$$

By Proposition 3.2 there exists a constant $\tau_\varepsilon > \tau_1$ such that

$$\| \theta(\cdot, \tau) \|_\infty \leqq 1 + \varepsilon \quad \text{for } \tau > \tau_\varepsilon. \tag{4.4}$$

Then it is easily seen that the function

$$\mathbf{z}(\tau) = -\left(\frac{1}{2} + \varepsilon + \frac{a_\varepsilon}{(\tau - \tau_1)}\right) < 0 \tag{4.5}$$

will be a subsolution of (4.3) in $\mathbf{R} \times (\tau_1, \infty)$, namely,

$$\mathbf{z}' \leqq 2\mathbf{z}^2 + (2\theta - 1)\mathbf{z} \quad \text{for } \tau > \tau_1 \tag{4.6}$$

if

$$\frac{a_\varepsilon(2a_\varepsilon - 1)}{(\tau - \tau_1)^2} - 2\mathbf{z}(1 + \varepsilon - \|\theta(\cdot, \tau)\|_\infty) + (1 + 2\varepsilon)\frac{a_\varepsilon}{\tau - \tau_1} \geqq 0 \tag{4.7}$$

for $\tau > \tau_1$. By using (4.4) we have that (4.7) holds for $\tau > \tau_\varepsilon$, and (4.7) will be also valid for $\tau_1 < \tau \leqq \tau_\varepsilon$ if $a_\varepsilon > 0$ is large enough. □

As a straightforward consequence of Proposition 3.2 and Lemma 4.1 we obtain a control on the derivative θ_x.

COROLLARY 4.2. *There exists a constant $C > 0$ such that*

$$|\theta_x(x, \tau)| \leqq C \quad \textit{in } \mathbf{R} \times (\tau_1, \infty). \tag{4.8}$$

5. Lower bound for the blow-up set and the asymptotic profile. We begin here the analysis of the asymptotic profile of the solution as $t \to T$. This analysis starts with a simple consequence of our semiconvexity result, Proposition 4.1.

LEMMA 5.1. *If $f \in \omega(\theta_0)$, then*

$$f_{xx} \geq -\tfrac{1}{2} \textit{ a.e. in } \mathbf{R}, \tag{5.1}$$

and, in particular,

$$f(x) \geqq \left(1 - \frac{x^2}{4}\right)_+ \quad \textit{a.e. in } \mathbf{R}. \tag{5.2}$$

Notice that, since (5.1) is true for every $f \in \omega(\theta_0)$, then necessarily

$$v(x,t) \geqq (T-t)^{-1}\left[\left(1 - \frac{x^2}{4}\right)_+ + o(1)\right]$$

in $\mathbf{R}$ as $t \to T$. This implies a first lower estimate of the blow-up set:

$$[-2, 2] \subset B, \quad \text{hence meas}(B) \geqq 4. \tag{5.3}$$

Although, according to the Introduction, these estimates will not be optimal in predicting the asymptotic shape, it is interesting to note that exactly (5.3) has been proved in [L] for the initial-boundary value problem by a completely different approach.

A sharp lower estimate is as follows.

THEOREM 5.2. *If $f \in \omega(\theta_0)$, then*

$$f(x) \geqq \cos^2\left(\frac{x}{2}\right) \quad \textit{for } |x| \leqq \pi, \tag{5.4}$$

so that

$$v(x,t) \geqq (T-t)^{-1}\left(\cos^2\left(\frac{x}{2}\right)+o(1)\right) \quad in\ \{|x| \leqq \pi\}. \tag{5.5}$$

It follows that $(-\pi,\pi) \subseteq B_*$, *so that* $\operatorname{meas}(B_*) \geqq 2\pi$.

Proof. Step 1. By the results of §3 we know that the maximum value of the variable $\theta(\cdot,t)$ tends to 1 as $t \to T$. By Lemma 5.1 we also have a first estimate from below for all elements of the ω-limit, namely,

$$f(0)=1,\ f_{xx} \geqq -\frac{1}{2} \quad \text{and} \quad f \geqq \left(1-\frac{x^2}{4}\right)_+ \quad \text{a. e. in } \mathbf{R}.$$

Step 2. *The Hamilton–Jacobi equation.* We next analyze the equation satisfied by the solution θ after some limit process. This turns out to be a Hamilton–Jacobi equation.

Fix an arbitrary sequence $\tau_j \to \infty$ such that $\theta(\cdot,\tau_j) \to f(\cdot) \in \omega(\theta_0)$ as $j \to \infty$. Using Propositions 3.2, 4.1, and Corollary 4.2 in passing to the limit as $\tau = \tau_j + s \to \infty$ in the linear and nonlinear terms of (1.12) we conclude that $\theta(\cdot,\tau_j+s) \to g(\cdot,s)$ as $j \to \infty$ locally in $L^\infty([0,\infty) : (c(\mathbf{R}))$, where $g(x,s)$ satisfies the following nonlinear Hamilton–Jacobi equation

$$g_s = (g_x)^2 + g^2 - g \quad \text{for } x \in \mathbf{R},\ s>0, \tag{5.6}$$

with initial data

$$g(x,0) = f(x) \quad \text{in } \mathbf{R}. \tag{5.7}$$

Step 3. *Explicit solutions.* Equation (5.6) with quadratic nonlinearities admits a family of explicit classical solutions, 2π-periodic in x, given by

$$g_*(x,s;\ \alpha) = \left(2\cos^2\left(\frac{x}{2}\right)+\alpha e^{-s}\right)(2+\alpha e^{-s})^{-1} \tag{5.8}$$

for $x \in \mathbf{R}$, $s>0$, where $\alpha > -2$ is a fixed constant. If $\alpha = 0$, then g_* is the classical stationary solution of (5.6):

$$g_*(x,s;\ 0) = \cos^2\left(\frac{x}{2}\right) \equiv G_*(x), \tag{5.9}$$

where $G_*(x) = G(x)$ for $|x| \leqq \pi$; see (1.16). For $\alpha > 0$ we have $g_*(x,s;\ \alpha) \geqq G_*(x)$ and

$$g_*(x,s;\ \alpha) \to G_*(x) \quad \text{as } s \to \infty \tag{5.10}$$

from above, uniformly in $\mathbf{R}$. On the other hand, it is easy to see that when $\alpha \in (-2,0)$ then $g_*(x,s;\ \alpha) = 0$ for $x = \pm x_*(s;\ \alpha)$ and $g_*(x,s;\ \alpha) > 0$ if $|x| < x_*(s;\ \alpha)$, where

$$x_*(s;\ \alpha) = 2\arccos(|\alpha|^{1/2}2^{-1/2}e^{-s/2}) < \pi$$

and $x_*(s;\ \alpha) \to \pi$ as $s \to \infty$. We also notice that if $\alpha \in (-2,0)$, then $g_*(x,s;\ \alpha) \leqq G_*(x)$ and

$$g_*(x,s;\ \alpha) \to G_*(x) \quad \text{as } s \to \infty \tag{5.11}$$

from below uniformly in $\mathbf{R}$.

Step 4. *Final estimate.* It follows from (5.2) that there exists $\alpha \in (-2, 0)$, which does not depend on f, such that

$$f(x) \geqq g_{-}(x, 0;\ \alpha) \quad \text{in } \mathbf{R}, \tag{5.12}$$

where

$$g_{-}(x, s;\ \alpha) = [g_*(x, s;\ \alpha)]_+ \quad \text{for } |x| \leqq \pi, \tag{5.13}$$

$$g_{-}(x, s;\ \alpha) = 0 \quad \text{for } |x| > \pi,\ s > 0, \tag{5.14}$$

is a viscosity solution of the Hamilton–Jacobi equation (5.6) (cf. [CL], [CEL]), and hence

$$g(x, s) \geqq g_{-}(x, s;\ \alpha) \quad \text{in } \mathbf{R} \times (0, \infty). \tag{5.15}$$

Then by (5.11) we have that for any small $\varepsilon > 0$ there exists $s_\varepsilon > 0$, which is independent for f, such that

$$g(x, s_\varepsilon) \geqq G(x) - \varepsilon \quad \text{for } |x| \leqq \pi \tag{5.16}$$

(recall that $G \equiv G_*$ in the interval $|x| \leqq \pi$). Since $\{\tau_j\}$ is an arbitrary sequence, we conclude that (5.16) holds for arbitrary $f \in \omega(\theta_0)$, whence (5.4) follows.

Remark. The lower estimate (5.4) can be also proved by an analysis similar to one given in §4. This can be done by using, instead of (4.3), the equation for the function $w = \theta_{xx} + \theta$. This equation has the form

$$w_\tau = e^{-\tau} w_{xx} + 2\theta_x w_x + (w - \theta)(2w - 1) + (\theta_x)^2 + \theta^2 - \theta. \tag{5.17}$$

To derive a lower bound of $w(x, \tau)$ as $\tau \to \infty$ we need a special estimate of the term $(\theta_x)^2 + \theta^2 - \theta$ in the right-hand side of (5.17).

6. Localization. In order to obtain upper bounds on the size of the blow-up set we shall make the additional assumption that v_0 has a finite number of intersections with an initial function of the form (2.6): $v_*(x, 0) = v_*(x, 0;\ T, \mu)$, with the same blow-up time $T > 0$ as u and $\mu \geqq \mu_T$ (which means that $v_*(x, 0)$ is nonnegative). It is easily seen that there exists a wide class of initial function with this property, in particular among the functions with compact support (see also Proposition 7.4 below). For a fixed $t \in [0, T)$ we denote by $N(t)$ the number of intersections of the solution $v(x, t)$ and the particular solution $v_*(x, t) = v_*(x, t;\ T, \mu) \geqq 0$ with the above initial data. By symmetry $N(t)$ is an even number. We then get the following estimate.

THEOREM 6.1. *Let T be the blow-up time of a solution u, and, with the above assumptions and notation, let*

$$N(0) = 2k < \infty. \tag{6.1}$$

Then the solution is uniformly bounded in set of the form

$$K = \{(x, t) : |x| \geqq x_1, 0 \leqq t < T\} \textit{ with } x_1 > L_k, \tag{6.2}$$

where $L_k = \pi(k + 2)$ if k is an odd number, and $L_k = \pi(k + 1)$ if k is even (L_k is a precise minimum for the function v_*). Consequently,

$$B \subseteq [-L_k, L_k]. \tag{6.3}$$

Proof. To begin with, we need the following result.

LEMMA 6.2. *Assume that as $t \to T$*

$$v(x,t) \geqq (T-t)^{-1}[1+\vartheta(t)] \tag{6.4}$$

for $|x| \leqq \pi/2$. Then

$$\vartheta(t) < (T-t)^2 \quad \text{as } t \to T. \tag{6.5}$$

Proof. By (3.24) we know that

$$T-t \leqq \frac{1}{\log E(t)}\left[1+\frac{1}{3}\frac{1}{\log^2 E(t)}(1+o(1))\right] \tag{6.6}$$

as $t \to T$. On the other hand, it follows from (6.4) and the definition of E, (3.18), that

$$E(t) \geqq \exp[(T-t)^{-1}(1+\vartheta(t))]-1 \quad \text{as } t \to T. \tag{6.7}$$

Then (6.5) is a direct consequence of (6.6) and (6.7). □

As a consequence of Lemma 6.2 and (2.8) we have the following consequence.

COROLLARY 6.3. *For $t \approx T$ there holds*

$$v(\pm 2\pi, t) < v_*(\pm 2\pi, t). \tag{6.8}$$

We continue the proof of Theorem 6.1 with the following.

LEMMA 6.4. *For $t \approx T$ we have*

$$v(\pm L_k, t) \leqq v_*(\pm L_k, t) \equiv \inf_{x\in\mathbf{R}} v_*(x,t). \tag{6.9}$$

Proof. Hypothesis (6.1) implies that

$$N(t) \leqq 2k \quad \text{for any } t \in (0,T). \tag{6.10}$$

By Corollary 6.3 we have that if $|x| \geqq 2\pi$ and $t \approx T$, then

$$v(x,t) < \sup_{x\in\mathbf{R}} v_*(x,t) \equiv v_*(\pm 2\pi, t). \tag{6.11}$$

We conclude that (6.9) holds. Indeed, if it is not valid and for $t \to T$,

$$v(\pm L_k, t) > v_*(\pm L_k, t), \tag{6.12}$$

then from (6.11) and (6.12) and using the spatial 2π-periodic structure of the explicit solution (2.1) we have the estimate $N(t) \geqq 2(k+1)$, contradicting (6.10). □

We are now ready to consider what happens in the domain $D = \{(x,t) : x > L_k, t_0 < t < T\}$ with $t_0 \approx T$. Observe first that since $v(|x|,t)$ is increasing in $|x|$, estimates (6.9) and (2.9) imply that for $t_0 \approx T$ we have

$$v(x,t) \leqq \tfrac{1}{2}|\log(T-t)|(1+o(1)) \tag{6.13}$$

in D. Consider the function $u(x,t) = e^{v(x,t)}-1$. It solves in D the problem consisting of (1.1) with initial condition

$$u(x,t_0) = a(x) \quad \text{for } x > L_k, \tag{6.14}$$

where a is bounded and nonnegative: $0 \leqq a(x) \leqq C$. It also satisfies a boundary condition for $t_0 \leqq t < T$ of the form

$$u(x_0, t) = h(t) \leqq \frac{1}{T-t} - 1. \tag{6.15}$$

Without loss of generality we may shift the axes and take $L_k = t_0 = 0$. The outer analysis of our solution relies on the following result.

LEMMA 6.5. *Let $u(x,t)$ be a solution of* (1.1) *in $D = \{(x,t) : x > 0, 0 < t < T\}$, which is nonincreasing as a function of x for all $t > 0$. If $u(x,0)$ is bounded for $x \geqq 0$, and there exists a constant $c > 0$ such that*

$$u(0,t) \leqq \frac{c}{T-t} - 1, \tag{6.16}$$

for $0 < t < T$, then $u(x,t)$ is bounded uniformly on sets of the form $K = \{(x,t) : x \geqq x_1, 0 \leqq t < T\}$ with $x_1 > 0$ arbitrary.

Proof. It proceeds via a two-step comparison argument. We obtain a first bound from above for $u(x,t)$ by comparing it with the solution $\overline{u}$ of the equation

$$\overline{u}_1 = \overline{u}_{xx} + \frac{c}{(T-t)} \log^2 \left(\frac{c}{T-t} \right) \tag{6.17}$$

with initial data C, a bound for $u(x,0)$ in $(0,\infty)$, and boundary data $c/(T-t) - 1$. Moreover, we may apply the superposition principle to the linear equation (6.17) and split $\overline{u}$ into $\overline{u}_1 + \overline{u}_2$, where $\overline{u}_1$ is a solution of (6.17) with zero initial and boundary data, and $\overline{u}_2$ solves the homogeneous heat equation $(\overline{u}_2)_t = (\overline{u}_2)_{xx}$ with same initial and boundary data as $\overline{u}$. By comparison with a solution that depends only on t, the function $\overline{u}_1$ can be easily estimated as

$$\overline{u}_1(x,t) \leqq \frac{c}{3} \left\{ \log^3 \left(\frac{c}{T-t} \right) - \log^3 \left(\frac{c}{T} \right) \right\}. \tag{6.18}$$

As for $\overline{u}_2$, by the results of [SGKM, p. 166] we can control the profile near the time T in the form

$$\overline{u}_2(x,t) \leqq C_1 + C_2 x^{-2} \quad \text{for some } C_1, C_2 > 0. \tag{6.19}$$

It follows that for $x \geqq x' > 0$ and as $t \to T$,

$$u(x,t) + 1 < c \log^3 \left(\frac{c}{T-t} \right). \tag{6.20}$$

The second step of the iteration consists in using (6.20) to modify (6.17) for the majorant into

$$\tilde{u}_t = \tilde{u}_{xx} + c \log^3 \left(\frac{c}{T-t} \right) \log^2 \left(c \log^3 \left(\frac{c}{T-t} \right) \right) \tag{6.21}$$

posed in $D' = \{(x,t) : x > x', 0 < t < T\}$ with corresponding initial and boundary conditions. We now get a finite estimate for both components of $\tilde{u}$ defined as before. Indeed, $\tilde{u}_1$ is uniformly bounded in D' since the last term in (6.21) is now integrable; $\tilde{u}_2$ is bounded for $x > x_1$ if $x_1 > x'$ by the same reason as in (6.19). □

With this proof of Theorem 6.1 is complete.

7. Minimal asymptotic behaviour. We are interested in giving a precise description of the shape of the solution at blow up. Under a stricter intersection condition on the data we establish in this section that the blow-up profile of our solution is just the minimal configuration (1.16). Our result is the following.

THEOREM 7.1. *Let as above v_0 be radially symmetric, nonincreasing in $|x|$, and assume moreover that it has two transversal intersections with a member $v_*(x, 0;\ T, \mu) \geqq 0$ of the explicit family* (2.1), *which has the same blow-up time T as v_0. Then as $t \to \infty$,*

$$\theta(x, \tau) \to G(x) = \begin{cases} \cos^2\left(\frac{x}{2}\right) & \textit{for } |x| \leqq \pi, \\ 0 & \textit{for } |x| > \pi, \end{cases} \tag{7.1}$$

uniformly in **R**.

Remark. We will give below in Proposition 7.4 direct conditions on v_0 so that this intersection hypothesis holds.

Proof. Step 1. *Perturbation analysis.* As explained in the Introduction the variable θ introduced in (1.11) satisfies (1.12):

$$\theta_\tau = e^{-\tau}\theta_{xx} + (\theta_x)^2 + \theta^2 - \theta$$

for $x \in \mathbf{R}$, $\tau > \tau_0 = -\log T$. This can be considered as a perturbation of the Hamilton–Jacobi equation

$$g_s = (g_x)^2 + g^2 - g \quad \text{for } x \in \mathbf{R},\ s > 0. \tag{HJ}$$

We want to describe the asymptotic behaviour of our solution of (1.12) in terms of the asymptotic behaviour of (HJ). More precisely, we want to apply Theorem 3 of [GV], which has to be conveniently adapted. [GV]'s result says that if three certain conditions to be discussed below are satisfied, then the ω-limit of the solution θ to (1.12) is contained in some part of the ω-limit of (HJ). In fact, we will need an improvement of that result and will show that, under our assumptions, such part is just the function G.

The first two steps in [GV] consist in proving that the orbits under consideration for (1.12) are compact in an appropriate topology and that, whenever a sequence $\theta(\cdot, \tau_j + s)$ converges locally in s to a function $g(\cdot, s)$ as $\tau_j \to \infty$, this g is a solution of (HJ). By the estimates proved in §4 we conclude that such facts are true and that the convergence takes place in $L^\infty_{\text{loc}}([0, \infty) : C(\mathbf{R}))$.

Step 2. *Stationary ω-limits for* (HJ). The last hypothesis in Theorem 3 of [GV] consists in showing the following:

(i) that the set Ω of all ω-limits of solutions g of (HJ) obtained in this way consists of stationary solutions; and

(ii) that it is nonvoid, compact and uniformly stable in the sense of Lyapunov for the flow generated by (HJ).

We first observe that the solutions of (HJ) we are considering take as initial data

$$g(x, 0) = f(x) \quad \text{in } \mathbf{R}, \tag{7.2}$$

where $f(\cdot) = \lim_{\tau_j \to \infty} \theta(\cdot, \tau_j)$ is an element of $\omega(\theta_0)$ satisfying

$$f = f(|x|) \quad \text{is nonincreasing in } |x|, \tag{7.3}$$

$$|f_x| \leqq C \text{ in } \mathbf{R}, \tag{7.4}$$

$$f(0) = 1, \qquad f_{xx} \geqq -\tfrac{1}{2} \quad \text{in } \mathbf{R}. \tag{7.5}$$

It is easily seen that the whole set of stationary and symmetric solutions $g = g(|x|)$ of (HJ), which are nonincreasing and tend to zero at infinity, is the one-parametric family of functions $S = \{\overline{g}(x;\ a), a \geqq 0\}$, where for a fixed $a \geqq 0$,

$$\tag{7.6} \begin{cases} \overline{g}(x, a) = 1 \quad \text{for } |x| \leqq a, \\ \overline{g}(x;\ a) = \cos^2\left(\dfrac{|x| - a}{2}\right) \quad \text{for } a \leqq |x| \leqq a + \pi, \\ \overline{g}(x;\ a) = 0 \quad \text{for } |x| \geqq a + \pi. \end{cases}$$

Nore that $G(x) \equiv \overline{g}(x;\ 0)$. We have the following.

LEMMA 7.2. *The ω-limit of g, solution of* (HJ), *with initial data f satisfying* (7.3)–(7.5), *consists of stationary solutions, i.e.*,

$$\tag{7.7} w_*(f) \subseteq S = \{\overline{g}(x;\ a), a \geqq 0\},$$

where ω_* denotes the ω-limit along the flow of (HJ).

Proof. We consider an arbitrary solution $g(x, s)$ of problem (HJ), (7.2), where the initial function $f(x)$ satisfies (7.3)–(7.5). Set

$$\tag{7.8} a = \sup\{x | f(x) = 1\} \geqq 0.$$

By Theorem 6.1 we have $a < \infty$, $f(x) = 1$ for $|x| \leqq a$ and $f(x) < 1$ for $|x| > a$. We shall compare $g(x, s)$ with the explicit solutions g_* of the Hamilton–Jacobi equation given in §5.

We now introduce a modification, g_a, of the function g_- given by (5.13), (5.14). It is defined for $a \geqq 0$ and $\alpha \in (-2, 0)$ as

$$\begin{aligned} &g_a(x, s;\ \alpha) = 1 && \text{for } |x| \leqq a, \\ &g_a(x, s;\ \alpha) = [g_*(|x| - a, s;\ \alpha)]_+ && \text{for } a \leqq |x| \leqq a + \pi, \\ &g_a(x, s;\ \alpha) = 0 && \text{for } |x| \geqq a + \pi, s \geqq 0. \end{aligned}$$

For any value of the parameter $\alpha \in (-2, 0)$, the function $g_a(x, t) = g_a(x, t;\ \alpha)$ thus defined is a subsolution of (HJ). By using estimates (7.5), which imply that

$$\tag{7.9} f(x) \geqq [1 - \tfrac{1}{4}(|x| - a)^2]_+ \quad \text{for } |x| \geqq a,$$

we deduce that there exists a value of the parameter $\alpha \in (-2, 0)$, which does not depend on f, such that, fixing this α,

$$\tag{7.10} f(x) \geqq g_a(x, 0) \quad \text{in } \mathbf{R}.$$

Then by comparison (cf. [CEL], [CL])

$$\tag{7.11} g(x, s) \geqq g_a(x, s) \quad \text{in } \mathbf{R} \times (0, \infty).$$

Similarly, we find a strictly positive super-solution g^a of (HJ) of the form

$$\begin{aligned} &g^a(x, s;\ \alpha, \lambda) = 1 && \text{for } |x| \leqq a + \lambda, \\ &g^a(x, s;\ \alpha, \lambda) = g_*(|x| - (a + \lambda), s;\ \alpha) && \text{for } |x| \geqq a + \lambda, s \geqq 0, \end{aligned}$$

with parameters α and $\lambda > 0$. One can see that for any arbitrarily small $\lambda > 0$ there exists $\alpha_+\lambda > 0$ such that

$$f(x) \leqq g^a(x,0) \quad \text{in } \mathbf{R},$$

and, therefore, by comparison

$$g(x,s) \leqq g^a(x,s) \quad \text{in } \mathbf{R} \times (0,\infty).$$

It follows from (5.10) and (5.11) that

$$g_a(x,s) \to \overline{g}(x;\, a), \qquad g^a(x,s) \to \overline{g}(x;\, a+\lambda) \tag{7.12}$$

as $s \to \infty$ uniformly in $\mathbf{R}$. Using the fact that $\overline{g}(x;\, a+\lambda) \to \overline{g}(x;\, a)$ as $\lambda \to 0$ uniformly in $\mathbf{R}$ completes the proof of the lemma. □

Step 3. *The reduced Ω-limit set.* The last condition we still have to check in order to apply Theorem 3 of [GV] is the uniform stability (in the sense of Lyapunov) of the Ω-limit set of (HJ). Unfortunately, such an assertion is false for the set S and the current topology. However, a careful inspection of the proof in [GV] show that we only work with a *reduced Ω-limit set,* which we define as "the set of ω-limits which actually occur as solutions of (HJ), (7.2) when the initial datum f is an ω-limit of (1.12)." Consequently, *we only have to prove that this reduced set has the desired properties.*

LEMMA 7.3. *The reduced Ω-limit set consists only of the special profile G.*

Proof. Let $v_*(x,t) = v_*(x,t;\, T,\mu)$ be the special solution referred to in the statement of the theorem. The proof of this lemma is based on a careful study of the comparison between our solution v and the special solution slightly delayed in time, $v_*(x,t+\varepsilon)$, $\varepsilon > 0$ small. We carefully control the relative situation of both solutions at $x = 0$ and $x = \pm\pi$. If an ω-limit is not G, then we show that the intersection count is violated as $t \to T$.

Let $N(t;\, \varepsilon)$ be the number of intersection in x at time $t \geqq 0$ of the solutions $v(x,t)$ and $v_*(x,t+\varepsilon)$ with initial data $v_0(x)$ and $v_*(x,\varepsilon)$. Our assumption is that $N(0;0) = 2$. By transversality and continuous dependence, the same is valid for the functions $v_0(x)$ and $v_*(x,\varepsilon)$ provided ε is small enough. This means that

$$N(0;\, \varepsilon) = 2 \quad \text{for any small } \varepsilon \geqq 0, \tag{7.13}$$

and hence by the Strong Maximum Principle we have

$$N(t;\, \varepsilon) \leqq 2 \quad \text{for } t \in [0, T-\varepsilon), \tag{7.14}$$

if $\varepsilon > 0$ is small enough. Notice that (7.14) with $\varepsilon = 0$ automatically implies the lower bound

$$v(0,t) > v_*(0,t) \quad \text{for } t \in [0,T). \tag{7.15}$$

Assume now that the assertion of the lemma is not true and that there exists a sequence $t_k \to T$ such that

$$\theta(x,t_k) \equiv (T-t_k)v(x,t_k) \to f(x)$$

as $k \to \infty$ uniformly in $\mathbf{R}$, and the function g, solution of (HJ) starting from f, tends in its turn as $s \to \infty$ to some $h \not\equiv G$. By Lemma 7.2 we have $h = \overline{g}(x;\, a)$ for some

$a > 0$. Therefore, we conclude from a simple triangular argument that there exists another sequence, that we denote by t_j, such that

$$\theta(x,t_j) \equiv (T-t_j)v(x,t_j) \to \overline{g}(x;\, a). \tag{7.16}$$

Notice that the corresponding explicit solution $v_*(x,t)$ satisfies (cf. (2.11))

$$\theta_*(x,t_j) \equiv (T-t_j)v_*(x,t_j) \to \overline{g}(x;\, 0) \tag{7.17}$$

as $j \to \infty$ uniformly in $\mathbf{R}$, and

$$\overline{g}(x;\, a) > \overline{g}(x;\, 0) \quad \text{for } 0 < |x| < a+\pi, \tag{7.18}$$

so that

$$\overline{g}(\pi;\, a) > 0 = \overline{g}(\pi;\, 0), \tag{7.19}$$

and

$$\overline{g}(0;\, a) = \overline{g}(0;\, 0) = 1. \tag{7.20}$$

Fix arbitrary j such that $s = T - t_j > 0$ is small enough and set $\varepsilon = s/2$. Then by (7.17), (7.16), and (7.20) we have that

$$\begin{aligned} v_*(0, t_j+\varepsilon) &\simeq (T-t_j-\varepsilon)^{-1} \equiv 2s^{-1} \\ &> s^{-1} \simeq v(0,t_j). \end{aligned} \tag{7.21}$$

The time displacement ε is used here. Using (7.17), (7.19), and (2.9) we get if $s > 0$ is small enough the inequality in the converse direction at $x = \pm\pi$;

$$\begin{aligned} v_*(\pm\pi, t_j+\varepsilon) &\simeq \frac{1}{2}|\log(T-t_j-\varepsilon)| \equiv \frac{1}{2}\left|\log\frac{s}{2}\right| \\ &< s^{-1}\overline{g}\,(\pm\pi;\, a) \simeq v(\pm\pi, t_j). \end{aligned} \tag{7.22}$$

One can see also that

$$v_*(\pm 2\pi, t_j+\varepsilon) > v(\pm 2\pi, t_j). \tag{7.23}$$

Estimates (7.21)–(7.23) imply that for the choice of $t = t_j$ and $\varepsilon > 0$ given above we have that

$$N(t_j;\, \varepsilon) \geqq 4,$$

which contradicts (7.14) for $t = t_j$. Hence $a = 0$, which completes the proof. □

Step 4. *Stability of G*. To end the proof we have to show that for every $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that if $g(\cdot, s)$ is a solution of (HJ), (7.2) such that

$$d(f, G) < \delta, \tag{7.24}$$

then

$$d(g(\cdot,s), G) < \varepsilon \quad \text{for every } s > 0 \tag{7.25}$$

(we denote by $d(\cdot,\cdot)$ the distance associated to $L^\infty(\mathbf{R})$). We can repeat the comparison arguments mentioned in Lemma 7.2, using the subsolution g_a and supersolution g_a. This completes the proof of Theorem 7.1.

To end the section we give explicit conditions under which the intersection assumption of Theorem 7.1 is fulfilled, as promised above.

PROPOSITION 7.4. *Let $v_0 = v_0(|x|)$ be a radially symmetric function, decreasing in $|x|$. Assume, moreover, that there exists a constant $m \in [0, v_0(0))$ such that*

$$v_0'(x) \leqq -\tfrac{1}{2}(v_0(0) - m)\sin x \tag{7.26}$$

for any $x \in [0,\pi]$ such that $v_0(x) > m$. Then the number of intersections of $v_0(x)$ with any positive function $v_*(x,0)$ of the form (2.6) *and satisfying*

$$v_0(0) > \max_x\{v_*(x,0)\} > \min_x\{v_*(x,0)\} \geqq m \tag{7.27}$$

is two and they are transversal. Moreover, if $v_0(0)$ is large and m small enough we can choose the parameters φ_0 and ψ_0 in (2.6) *so that $v_*(x,t)$ has the same blow-up time as v, and* (7.26) *holds.*

Proof. The initial function $\overline{v}_*(x) \equiv v_*(x,0;\,T,\mu)$ has the form (2.6) for $\mu \geqq \mu_T$, i.e.,

$$\overline{v}_*(x) = \varphi_0(\psi_0 + \cos x), \tag{7.28}$$

where $\varphi_0 > 0$, $\psi_0 \geqq 1$ are constants. According to (7.27),

$$M_1 = \inf_x \overline{v}_*(x) \equiv \varphi_0(\psi_0 - 1) \geqq m, \tag{7.29}$$

and

$$M_2 = \sup_x \overline{v}_*(x) \equiv \varphi_0(\psi_0 + 1) < v_0(0), \tag{7.30}$$

where m is as in (7.26). One can see that under hypothesis (7.26) the function $v_0(x)$ intersects an arbitrary function of the family $\overline{v}_*(x)$ exactly at two points. In fact, at any intersection point $x > 0$ we have

$$\overline{v}_*'(x) = -\varphi_0 \sin x > v_0'(x),$$

since $v_0(0) - m > M_2 - M_1 = 2\varphi_0$.

Finally, in order for v_* to have the same blow-up time T as u, and according to the analysis of §2, the data φ_0 and ψ_0 have to satisfy a certain condition, namely, the point (φ_0, ψ_0) must lie on some curve situated in the region $\{0 < \varphi_0 < \infty, 1 < \psi_0 < \infty\}$. This condition is clearly compatible with (7.29), (7.30) if m is small and $v_0(0)$ large. Then there are infinitely many choices. □

8. Minimal blow-up set. We have established minimal asymptotic behaviour in terms of the variable θ in the preceding section. Unfortunately, due to the factor $(T-t)$ in the change of variables (1.11), this does not automatically imply that u remains bounded outside of the minimal set $[-\pi,\pi]$. This section is devoted to establishing such a result.

THEOREM 8.1. *Assume that v_0 satisfies the conditions of Proposition* 7.4. *Then $B \equiv [-\pi,\pi]$.*

Remark. In view of Theorem 5.2, it follows that meas $(B_*) = 2\pi$, so that the solution actually converges to infinity as $t \to T$ at any point of the interval $-\pi < x < \pi$; see (8.1) below. The actual asymptotic behaviour at $x = \pm\pi$ is an open question.

Proof. The proof consists of a preliminary analysis of what happens near $x = \pm\pi$ (*boundary analysis*) plus an analysis of the situation for $|x| > \pi$. This latter outer analysis follows Lemma 6.5. To begin with, by Theorem 7.1 we know that as $t \to T$,

$$\theta(x,t) \equiv (T-t)v(x,t) \to G(x) \tag{8.1}$$

uniformly in **R**. The special solution with the same blow-up time selected as in Proposition 7.4 satisfies

$$\theta_*(x,t) \equiv (T-t)v_*(x,t) \to G_*(x) \tag{8.2}$$

as $t \to T$ uniformly in **R**. For a fixed $t \in [0,T)$ and arbitrary $\lambda \geqq 0$ small enough we denote by $N(t;\ \lambda)$ the number of intersections in x of the solutions $v(x,t)$ and this $v_*(x-\lambda,t)$ having the same blow-up time T as v. Then, since by Proposition 7.4 each intersection of the initial functions $v_0(x)$ and $v_*(x,0)$ is transversal, by continuity of the function $v_*(x-\lambda,0)$ with respect to λ we conclude that there holds (cf. (7.13))

$$N(0;\ \lambda) = 2 \quad \text{for any small } \lambda \geqq 0. \tag{8.3}$$

Hence (cf. (7.14))

$$N(t;\ \lambda) \leqq 2 \quad \text{for } t \in [0,T), \tag{8.4}$$

if $\lambda \geqq 0$ is small enough (notice that (7.26) yields that (8.4) is valid for arbitrary $\lambda \geqq 0$).

We now perform the boundary analysis. We prove that the solution $v(x,t)$ is small enough as $t \to T$ in a right neighborhood of the point $x = \pi$. (By symmetry, the same is true for a left neighborhood of $x = -\pi$).

LEMMA 8.2. *Let* $\lambda_0 > 0$ *be small enough. Then, as* $t \to T$,

$$v(\pi+\lambda_0,t) \leqq v_*(\pi,t) \equiv \tfrac{1}{2}|\log(T-t)|(1+o(1)). \tag{8.5}$$

Proof. We use an interesting technique of intersection comparison with shifting in x. Let $x = x_0(t) > 0$ be the unique positive intersection point of the solutions $v(x,t)$ and $v_*(x,t)$. Notice that since both solutions are analytic functions in the x variable, each intersection point is isolated for $0 < t < T$, [Fr2]. If $x_0(t) \leqq \pi$ as $t \to T$, then using (7.15) and (8.4) with $\lambda = 0$ we deduce that $v(x;\ t) \leqq v_*(x,t)$ for $x \geqq \pi$, and hence (8.5) is valid even with $\lambda_0 = 0$.

Suppose on the contrary that there exists a monotone sequence $t_j \to T$, such that

$$x_0(t_j) > \pi \quad \text{for } j = 1,2,\ldots. \tag{8.6}$$

Then by the Strong Maximum Principle [Fr1] we have that

$$v(x,t_j) > v_*(x,t_j) \quad \text{for } |x| \leqq \pi. \tag{8.7}$$

Hence,

$$\delta_j = \delta(t_j) \equiv \sup\{\lambda > 0 : v(x,t_j) \geqq v_*(x-\lambda,t_j) \text{ for } |x| \leqq \pi\} > 0. \tag{8.8}$$

It follows from (8.1) and (8.2) that $\delta_j \to 0$ as $j \to \infty$.

Compare now for a fixed j large enough the functions $v(x,t_j)$ and $v_*(x-\delta_j,t_j)$. First, we conclude that

$$v(x,t_j) > v_*(x-\delta_j,t_j) \quad \text{for } x \in [0,\pi). \tag{8.9}$$

Otherwise, if (8.9) is not valid so that, by the definition of δ_j, there exists some interior tangency point $x = x_1 \in (0, \pi)$ of these functions, at which

$$\begin{aligned} v(x_1, t_j) &= v_*(x_1 - \delta_j, t_j), \\ v_x(x_1, t_j) &= (v_*)_x(x_1 - \delta_j, t_j); \end{aligned} \tag{8.10}$$

we arrive at the contradiction with (8.4). Indeed, under conditions (8.10) by additionally "shifting to the right in x" the function $v_*(x - \delta_j, t_j)$, we deduce that for arbitrary $0 < \nu \ll \delta_j$ the number of intersections grows,

$$N(t_j;\, \delta_j + \nu) \geqq 3, \tag{8.11}$$

contradicting (8.4) with $t = t_j$, $\lambda = \delta_j + \nu$.

Thus, (8.9) holds. From the definition of δ_j given by (8.8) we conclude then that necessarily

$$v_*(\pi - \delta_j, t_j) = v(\pi, t_j), \tag{8.12}$$

and, therefore, (8.4) implies that

$$v(\pi + \delta_j, t_j) < v_*(\pi, t_j). \tag{8.13}$$

Since $\{t_j\}$ is arbitrary, by using the fact that $\delta(t) \to 0$ as $t \to T$ we have that for any small $\lambda_0 > 0$ there exists $t_0 \in (0, T)$ such that (8.5) holds for $t \in (t_0, T)$. □

It then follows from (2.9) that for $t \to T$,

$$v(x_0, t) \leqq -\tfrac{1}{2}\log(T - t)(1 + o(1)), \tag{8.14}$$

where $x_0 = \pi + \lambda_0$. The proof that u is uniformly bounded on sets of the form $K = \{(x, t) : x \geqq x_1, 0 \leqq t < T\}$ with $x_1 > x_0$ is done by applying Lemma 6.5. Since $\lambda > 0$ is arbitrarily small, the proof of Theorem 9.1 is thus complete. □

9. Periodic solutions. The techniques developed in the preceding sections can be easily adapted to some other situations. We consider in this section the case of periodic initial data with period a multiple of 2π, namely,

$$u_0(x + 2m\pi) = u_0(x), \tag{9.1}$$

with m and integer. Let us take the interval $I_m = [-m\pi, m\pi]$ as the basic period. We also assume that u_0 is symmetric in I_m, nonnegative, and nonincreasing for $0 \leq x \leq m\pi$. We also ask u_0 not to be constant in order to avoid trivial cases whose behaviour is different from the one we want to study here. Notice that we can think of the solution of the Cauchy Problem for (1.1) in $Q = \mathbf{R} \times (0, T)$ under these initial data as the solution of the Neumann Problem posed in $Q_m = I_m \times (0, T)$ with boundary data

$$u_x(\pm m\pi, t) = 0 \quad \text{for } t > 0. \tag{9.2}$$

To begin with, the explicit solutions u_* considered in §2 are still solutions of this problem. Moreover, for any solution of the Neumann Problem we have the following:

(i) If t is fixed the solution $u(x, t)$ is a symmetric function of x, decreasing for $0 < x < m\pi$. Hence, the maximum is always taken at $x = 0$. The Maximum Principle applies and the intersection number in I_m is still a nonincreasing function of time;

(ii) Estimates (3.1) and (3.4) of §3 hold. The proofs are unchanged;

(iii) The same is true for semiconvexity results of §4 and the minimal profiles of §5;

(iv) Theorem 6.1 and the other results of §6 are valid, though statements (6.2), (6.3) are void if $m\pi \leqq L_k$.

Thus, we obtain the next theorem.

THEOREM 9.1. *Let u be a solution of* (1.1), (9.1), (9.2) *with* $m \geqq 4$. *Assume that u_0 satisfies the intersection condition $N(0) = 2$ with some nonnegative initial data $u_*(x, 0;\ T, \mu)$, $\mu \geqq \mu_T$. Then*

$$B \cap I_m \subseteq [-3\pi, 3\pi]. \tag{9.3}$$

Proof. Our assumption is that $v_0 = \log(1 + u_0)$ satisfies

$$N(0) = 2, \tag{9.4}$$

where $N(t)$ denotes number of intersections in I_m of the functions $v(x, t)$ and $v_*(x, t)$. As before it follows that

$$N(t) \leqq 2 \tag{9.5}$$

for every $0 < t < T$. As in Corollary 6.3 we have $v(\pm 2\pi, t) < v_*(\pm 2\pi, t)$ for $t \approx T$. This and (9.5) imply that

$$v(3\pi, t) < v_*(3\pi, t) = \tfrac{1}{2}|\log(T - t)|(1 + o(1)) \tag{9.6}$$

as $t \to T$. Since the solution $v(x, t)$ is nonincreasing in $x > 0$ we conclude that for $t \approx T$ and $3\pi \leqq x \leqq m\pi$,

$$v(x, t) \leqq |\log(T - t)|. \tag{9.7}$$

We now apply Lemma 6.5 to conclude that v is uniformly bounded for $m\pi \geqq x \geqq x_1 > 3\pi$ and $0 < t < T$. □

(v) Under stricter conditions on the initial data we obtain, as in Theorem 7.1, the following.

THEOREM 9.2. *Let $m \geqq 3$ and let $v_0 = \log(1 + u_0)$ satisfy the conditions of Proposition 7.4 in I_m. Then*

$$B \cap I_m = [-\pi, \pi]. \tag{9.8}$$

Proof. We have (9.4) and then (9.5). Arguments similar to those given in §§7 and 8 show that, similar to (9.6) in the preceding result, the inequality $v(\pi, t) \leqq v_*(\pi, t)$ is valid after slightly shifting in x and t the explicit solution $v_*(x, t)$. In this way we can apply the arguments of Theorems 7.1 and 8.1. □

Acknowledgments. The first author thanks the Departamento de Matemáticas and the Universidad Autónoma de Madrid for their hospitality during his visits to work in this project. The second author is similarly grateful to the Keldysh Institute of Applied Mathematics and the Russian Academy of Sciences. This paper was completed at the Institute for Mathematics and its Applications, University of Minnesota.

REFERENCES

[A] S. ANGENENT, *The zero set of a solution of a parabolic equations*, J. Reine Angew. Math., 390 (1988), pp. 79–96.

[BE] J. Bebernes and D. Eberly, *Mathematical Problems from Combustion Theory*, Applied Math. Sciences, Vol. 83, Springer-Verlag, New York, 1989.

[CM] X. Y. Chen and H. Matano, *Convergence, asymptotic periodicity and finite point blow-up in one-dimensional semilinear heat equations*, J. Differential Equations, 78 (1988), pp. 160–190.

[CEL] M. G. Crandall, L. C. Evans, and P. L. Lions, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.

[CL] M. G. Crandall and P. L. Lions, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 272 (1983), pp. 1–42.

[Fm] A. Friedman and B. McLeod, *Blow-up of positive solutions of semilinear heat equation*, Indiana Univ. Math. J., 34 (1985), pp. 425–447.

[Fr1] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[Fr2] ———, *On the regularity of the solutions of non-linear elliptic and parabolic systems of partial differential equations*, J. Math. Mech., 7 (1958), pp. 43–59.

[Fu] H. Fujita, *On some nonexistence and nonuniqueness theorems for nonlinear parabolic equations*, Proc. Symp. in Pure Math. 18, American Mathematical Society, Providence, RI, 1969, 105–113.

[G1] V. A. Galaktionov, *On global nonexistence to Cauchy problems for quasilinear parabolic equations*, Zh. Vychisl. Mat. i Mat. Fiz., 23 (1983), pp. 1072–1087. (In Russian.)

[G2] ———, *On blow-up and degeneracy for the semilinear heat equation with source*, Proc. Roy. Soc. Edinburgh Sect. A, 115A (1990), pp. 19–24.

[G3] ———, *On a blow-up set for the quasilinear heat equation* $u_t = (u^\sigma u_x)_x + u^{\sigma+1}$, J. Differential Equations, 101 (1993), pp. 66–79.

[GKMS] V. A. Galaktionov, S. P. Kurdyumov, A. P. Mikhailnov, and A. A. Samarskii, *On unbounded solutions of semilinear parabolic equations*, Keldysh Inst. Appl. Math. Acad. Sci. USSR #161, 1979, preprint. (In Russian.)

[GKS1] V. A. Galaktionov, S. P. Kurdyumov, and A. A. Samarskii, *On approximate self-similar solutions for some class of quasilinear heat equations with sources*, Mat. Sb., 124 (1984), pp. 163–188. (In Russian.)

[GKS2] ———, *On the method of stationary states for quasilinear parabolic equations*, Mat. Sb., 180 (1989) (in Russian); Math. USSR Sbornik, 67 (1990), pp. 449–471. (In English.)

[GP1] V. A. Galaktionov and S. A. Posashkov, *Applications of new comparison theorems for unbounded solutions of a nonlinear parabolic equation*, Differentsial'nye Uravneniya, 22 (1986), pp. 1165–1173. (In Russian.)

[GP2] ———, *On some properties of evolution of unbounded solutions to semilinear parabolic equations*, Keldysh Inst. Appl. Math. Acad. Sci. URSS, 232, (1987), preprint. (In Russian.)

[GP3] ———, *On new explicit solutions of parabolic equations with quadratic nonlinearities*, Zh. Vychisl. Mat. i Mat. Fiz., 29 (1989), pp. 497–506. (In Russian.)

[GP4] ———, *Any large solution of nonlinear heat conduction equation becomes monotone in time*, Proc. Roy. Soc. Edinburgh Sect. A., 118A (1991), pp. 13–20.

[GV] V. A. Galaktionov and J. L. Vazquez, *Asymptotic behaviour of nonlinear parabolic equations with critical exponents. A dynamical systems approach*, J. Funct. Anal., 100 (1991), pp. 435–462.

[Kal] A. S. Kalashnikov, *Some questions of the qualitative theory of nonlinear degenerate parabolic second-order equations*, Uspekhi Mat. Nauk., 42 (1987), pp. 135–176. (In Russian.)

[K] S. Kaplan, *On the growth of solutions of quasilinear parabolic equations*, Comm. Pure Appl. Math., 16 (1963), pp. 305–330.

[L] A. A. Lacey, *Global blow-up of a nonlinear heat equation*, Proc. Roy. Soc. Edinburgh Sect. A., 104A (1986), pp. 161–167.

[M] H. Matano, *Nonincrease of the lap number of a solution for a one-dimensional semi-linear parabolic equation*, J. Fac. Sci. Univ. Tokyo, Sect. IA Math., 29 (1982), pp. 401–441.

[Sam] A. A. Samarskii, *On new investigation methods of asymptotic properties of parabolic equations*, Trudy Mat. Inst. Steklov. USSR, Vol. 158, 1981, pp. 153–16. (In Russian.)

[SGKM] A. A. Samarskii, V. A. Galaktionov, S. P. Kurdyumov, and A. P. Mikhailnov, *Blow-up in Problems for Quasilinear Parabolic Equations*, Nauka, Moscow, 1987. (In Russian; English translation: Walter de Gruyter, Berlin, to appear.)

[Sat] D. H. Sattinger, *On the total variation of solutions of parabolic equations*, Math. Ann., 183 (1969), pp. 78–92.

[V] J. J. L. Velazquez, *Estimates on* $(N-1)$*-dimensional Hausdorff measure of the blow-up set for a semilinear heat equation*, IMA Preprint Series, 964, April 1992.

SIAM J. MATH. ANAL.
Vol. 24, No. 5, pp. 1277–1282, September 1993

008

ON THE NODAL SETS OF THE EIGENFUNCTIONS OF CERTAIN HOMOGENEOUS AND NONHOMOGENEOUS MEMBRANES*

CHAO-LIANG SHEN†

Abstract. The author finds conditions for the existence of eigenfunctions whose nodal sets consist of one line segment or two perpendicular line segments. Using nodal sets of eigenfunctions, a sufficient condition is found for the density function of a nonhomogeneous annular membrane that is a radial function. The author also shows how to use the nodal sets of eigenfunctions of some particular type and the corresponding eigenvalues to determine the density function of the annular membrane if its density function is a radial function.

Key words. nonhomogeneous membranes, density functions, eigenvalues, eigenfunctions, nodal sets, nodal domains

AMS subject classifications. 35B05, 35P99

1. Introduction. Suppose Ω is a simply connected plane domain which is symmetric with respect to the x-axis. Does the fixed membrane problem

$$\Delta\varphi + \lambda\varphi = 0 \quad \text{in } \Omega, \qquad \varphi = 0 \quad \text{on } \partial\Omega \tag{1.1}$$

have an eigenfunction whose nodal set is the portion of x-axis in Ω? If Ω is also symmetric with respect to the y-axis, does (1.1) have an eigenfunction whose nodal set consists of the portion of x-axis and the portion of y-axis in the domain Ω? These questions shall be answered in §2.

In §3 we consider the nonhomogeneous membrane problem over an annulus R in $\mathbb{R}^2$:

$$\Delta\varphi + \lambda\rho(x,y)\varphi = 0 \quad \text{in } R, \qquad \varphi = 0 \quad \text{on } \partial\Omega, \tag{1.2}$$

where the density function $\rho(x,y)$ is a positive continuous function. We are interested in the question of determining the density function $\rho(x,y)$ by knowing the nodal sets of sufficiently many eigenfunctions and the corresponding eigenvalues. We prove that if (1.2) has infinitely many nodal sets of the eigenfunctions that are *circular-nested*, then the density function $\rho(x,y)$ must be radial (Theorem 3.1). Furthermore, if the density function ρ in (1.2) is radial, then, using some results in [5], we show that sufficiently many *circular nodal sets* are enough to determine the density function ρ (Theorem 3.3).

2. Existence of eigenfunctions whose nodal sets consist of straight lines. Let $f(x)$ be a continuous function on the interval $a \leqq x \leqq b$, $f(a) = f(b) = 0$, $f(x) > 0$ for $a < x < b$. We shall use the notation Ω_f to denote the region $\{(x,y)\} \in \mathbb{R}^2 : a < x < b,\ -f(x) < y < f(x)\}$. In this paper we shall assume f is nice enough for Ω_f to have smooth boundary. We use the notation $N(\varphi)$ to denote the *nodal set* of an eigenfunction φ of (1.1) on Ω, which is the set $\{(x,y) \in \Omega : \varphi(x,y) = 0\}$. The notation $E_4(\Omega_f)$ is used to denote the four-dimensional domain whose closure is

$$\overline{E_4(\Omega_f)} = \{(x, y_1, y_2, y_3) \in \mathbb{R}^4 : a \leqq x \leqq b, y_1^2 + y_2^2 + y_3^2 \leqq [f(x)]^2\}. \tag{2.1}$$

If the even function $g(x)$ is continuous on the interval $-a \leqq x \leqq a$, strictly positive in the interior, and vanishes at $\pm a$, we shall use the notation $E_6(\Omega_g)$ to denote the six-

* Received by the editors December 24, 1991; accepted for publication (in revised form) December 2, 1992. This research was supported in part by the National Science Council of the Republic of China.

† Institute of Mathematics, National Tsing Hua University, Hsinchu, Taiwan 30043, Republic of China.

dimensional domain whose closure is defined as follows:

$$\overline{E_6(\Omega_g)} = \{(x_1, x_2, x_3; y_1, y_2, y_3) \in \mathbb{R}^6 : x_1^2 + x_2^2 + x_3^2 \leqq a^2, \\ y_1^2 + y_2^2 + y_3^2 \leqq [g((x_1^2 + x_2^2 + x_3^2)^{1/2})]^2 \}. \tag{2.2}$$

We shall use the notation $\Delta_{x;y_1,y_2,y_3}$, $\Delta_{x_1,x_2,x_3;y_1,y_2,y_3}$ to denote the four-dimensional and six-dimensional Laplacians, respectively:

$$\Delta_{x;y_1,y_2,y_3} = \frac{\partial^2}{\partial x^2} + \sum_{j=1}^{3} \frac{\partial^2}{\partial y_j^2},$$

$$\Delta_{x_1,x_2,x_3;y_1,y_2,y_3} = \sum_{j=1}^{3} \frac{\partial^2}{\partial x_j^2} + \sum_{j=1}^{3} \frac{\partial^2}{\partial y_j^2}.$$

The following simple results shall be used later. Since they follow easily from calculation, we omit the proof.

LEMMA 2.1. (1) *Suppose the function $v(x; y_1, y_2, y_3)$ depends only on x and $y_1^2 + y_2^2 + y_3^2$. Then*

$$\Delta_{x;y_1,y_2,y_3} v = \Delta_{x,y} v(x, y) + \frac{2v_y(x, y)}{y}, \tag{2.3}$$

where $y^2 = y_1^2 + y_2^2 + y_3^2$, $v(x, y) = v(x; y_1, y_2, y_3)$, $v_y = \partial v/\partial y$, and $\Delta_{x,y}$ is the two-dimensional Laplacian in x, y.

(2) *Suppose the function $w(x_1, x_2, x_3; y_1, y_2, y_3)$ depends only on $x^2 = x_1^2 + x_2^2 + x_3^2$ and $y^2 := y_1^2 + y_2^2 + y_3^2$. Then*

$$\Delta_{x_1,x_2,x_3;y_1,y_2,y_3} w = \Delta_{x,y} w(x, y) + \frac{2w_x}{x} + \frac{2w_y}{y}, \tag{2.4}$$

where $w(x, y) = w(x_1, x_2, x_3; y_1, y_2, y_3)$.

In [4], the author had used a method to compare two eigenvalues by knowing the nodal sets of the corresponding eigenfunctions. One of the key points of the method is to construct a four-dimensional domain from Ω_f by revolving Ω_f about the x-axis. We use the same idea to prove the following theorem.

THEOREM 2.2. *For the region Ω_f, the eigenvalue problem* (1.1) *has an eigenfunction φ whose nodal set is the portion of x-axis in Ω_f, i.e., $N(\varphi) = \{(x, 0) : a < x < b\}$.*

Proof. On the four-dimensional domain $E_4(\Omega_f)$ we consider the following eigenvalue problem:

$$\Delta_{x;y_1,y_2,y_3} v + \mu v = 0 \quad \text{in } E_4(\Omega_f), \qquad v = 0 \quad \text{on } \partial E_4(\Omega_f). \tag{2.5}$$

Let v_1 be the first eigenfunction of the eigenvalue problem (2.5) such that $v_1(0; 0, 0, 0) = 1$. Using the fact that the first eigenvalue of (2.5) is simple, and the fact that $E_4(\Omega_f)$ is axially symmetric, we see that for any three-dimensional rotation $A = [a_{ij}]_{i,j=1}^{3}$, $v_1(x; y_1, y_2, y_3) = v_1(x; \sum_{j=1}^{3} a_{1j} y_j, \sum_{j=1}^{3} a_{2j} y_j, \sum_{j=1}^{3} a_{3j} y_j)$. This implies that v_1 depends only on x and $y_1^2 + y_2^2 + y_3^2$. Let $y^2 = y_1^2 + y_2^2 + y_3^2$. Again we use the notation $v_1(x, y)$ to denote $v_1(x; y_1, y_2, y_3)$. Then Lemma 2.1 tells us that

$$-\Delta_{x,y} v_1 - \frac{2(v_1)_y}{y} = \mu_1 v_1. \tag{2.6}$$

Let $\varphi(x,y) = yv_1(x,y)$. Then it is clear that $\varphi(x,y) = 0$ on $\partial\Omega_f$, $\varphi(x,0) = 0$ for $a < x < b$. Furthermore, by (2.6),

$$\begin{aligned}\Delta\varphi &= y\Delta_{x,y}v_1 + 2(v_1)_y \\ &= -\mu_1 y v_1 \\ &= -\mu_1\varphi,\end{aligned}$$

i. e., μ_1 is an eigenvalue and φ is an eigenfunction of (1.1), which has the stated property. □

Following a similar idea we have the next theorem.

THEOREM 2.3. *Let $g(x)$ be an even function which is continuous on $-a \leqq x \leqq a$, $g(a) = g(-a) = 0$, and $g(x) > 0$ for all $-a < x < a$. Then, for Ω_g, the eigenvalue problem* (1.1) *has an eigenfunction ψ whose nodal set $N(\psi) = \{(x,0) : -a < x < a\} \cup \{(0,y) : -g(0) < y < g(0)\}$.*

Proof. On $E_6(\Omega_g)$ we consider the following eigenvalue problem:

$$\Delta_{x_1,x_2,x_3;y_1,y_2,y_3}w + vw = 0 \quad \text{in } E_6(\Omega_g), w = 0 \quad \text{on } \partial E_6(\Omega_g). \tag{2.7}$$

Let v_1, w_1 denote the first eigenvalue and the first eigenfunction with $w_1(0,0,0;0,0,0) = 1$, respectively. Then, using the simplicity of v_1, we have

$$\begin{aligned}&w_1\left(\textstyle\sum_j a_{1j}x_j, \sum_j a_{2j}x_j, \sum_j a_{3j}x_j; \sum_j b_{1j}x_j, \sum_j b_{2j}x_j, \sum_j b_{3j}x_j\right) \\ &\quad = w_1(x_1,x_2,x_3;y_1,y_2,y_3)\end{aligned}$$

for all three-dimensional rotations $[a_{ij}]$, $[b_{ij}]$. This implies that $w_1(x_1,x_2,x_3;y_1,y_2,y_3)$ depends only on $x^2 = x_1^2 + x_2^2 + x_3^2$, and $y^2 = y_1^2 + y_2^2 + y_3^2$. Denote this w_1 by $w_1(x,y)$. Then, by Lemma 2.1 we have

$$-\Delta_{x,y}w_1 - \frac{2(w_1)_x}{x} - \frac{2(w_1)_y}{y} = v_1 w_1. \tag{2.8}$$

Now we define $\psi(x,y) = xyw_1(x,y)$. Then

$$\begin{aligned}\psi(x,y) &= 0 &&\text{on } \partial\Omega_g, \\ \psi(x,0) &= 0 &&\text{for } -a < x < a, \\ \psi(0,y) &= 0 &&\text{for } -g(0) < y < g(0).\end{aligned}$$

Furthermore, by (2.8) we have

$$\begin{aligned}\Delta_{x,y}\psi &= xy\Delta_{x,y}w_1 + 2x(w_1)_y + 2y(w_1)_x \\ &= -v_1 xyw_1 \\ &= -v_1\psi.\end{aligned}$$

Thus ψ is the desired eigenfunction. □

3. Nonhomogeneous annular membranes with radial density functions. Let R be the annulus $r_0^2 < x^2 + y^2 < r_1^2$, $0 < r_0 < r_1$. Suppose (λ, φ) is an eigenpair of (1.2) such that the nodal set of φ consists of concentric circles centered at the origin $(0,0)$. We shall call the nodal set a *circular nodal set*, φ a *circular eigenfunction*, and λ a *circular eigenvalue*.

If $(\tilde{\lambda}, \tilde{\varphi})$ is an eigenpair of (1.2) such that the nodal set of $\tilde{\varphi}$ is of the form $\{(r\cos\theta, \sin\theta)$: $r = r_{1,\tilde{\varphi}}, \ldots, r_{l(\tilde{\varphi}),\tilde{\varphi}}$, or $r_0 < r < r_1$ and $\theta = \theta_* + k\pi/m(\tilde{\varphi}), k = 1, 2, \ldots, 2m(\tilde{\varphi})\}$, where $r_0 < r_{1,\tilde{\varphi}} < \cdots < r_{l(\tilde{\varphi}),\tilde{\varphi}} < r_1$, θ_* and the integers $l(\tilde{\varphi})$, $m(\tilde{\varphi})$ are determined by the eigenfunction $\tilde{\varphi}$, then we call such an eigenfunction $\tilde{\varphi}$ a *circular-nested eigenfunction*, and the corresponding eigenvalue $\tilde{\lambda}$ a *circular-nested eigenvalue*. For an eigenfunction φ of (1.2), we use mesh (φ) to denote the maximum of the diameters of its nodal domains. Throughout this section we shall assume the density function $\rho(x, y)$ in (1.2) is a positive continuous function. We have the following inverse spectral theorem.

THEOREM 3.1. *If* (1.2) *has infinitely many circular-nested eigenfunction* φ_{n_j} *such that* $\lim_{n_j\to\chi}$ *mesh* $(\varphi_{n_j}) = 0$, *then the density function* $\rho(x, y)$ *is a radial function, i.e.*, $\rho(x, y) = \rho(r), r = (x^2 + y^2)^{1/2}$.

Proof. By assumption, the nodal domains of φ_{n_j} are of the form (in polar coordinate):

$$I^{(n_j)}(i,k) = \left\{(r,\theta) : r_i^{(n_j)} < r < r_{i+1}^{(n_j)}, \theta_{n_j} + \frac{k\pi}{m(n_j)} < \theta < \theta_{n_j} + \frac{(k+1)\pi}{m(n_j)}\right\},$$

where $r_0 = r_0^{(n_j)} < r_1^{(n_j)} < r_{l(nj)}^{(n_j)} = r_1$, $k = 0, 1, \ldots, 2m(n_j) - 1$; the integers $l(n_j)$, $m(n_j)$, the angle θ_{n_j}, and the radii $r_i^{(n_j)}$ are determined by φ_{n_j}.

If the density function ρ is not radial, then there exist an r_*, $r_0 < r_* < r_1$, and θ_*, θ_{**} such that $\rho(r_*\cos\theta_*, r_*\sin\theta_*) > \rho(r_*\cos\theta_{**}, r_*\sin\theta_{**})$. Since $\lim_{\lambda_{n_j}\to\infty}$ mesh $(\varphi_{n_j}) = 0$, there exist indices n_j, i, k_1, k_2, $k_1 \neq k_2$, such that (r_*, θ_*) is in closure of $I^{(n_j)}(i, k_1)$, (r_*, θ_{**}) is in the closure of $I^{(n_j)}(i, k_2)$, and

$$\rho(x_1, y_1) > \rho(x_2, y_2) \quad \text{for all } (x_1, y_1) \text{ in } I^{(n_j)}(i, k_1), \quad \text{all } (x_2, y_2) \text{ in } I^{(n_j)}(i, k_2). \tag{3.1}$$

Let $\lambda_1(I^{(n_j)}(i, k_s))$ be the first eigenvalue of

$$\Delta u + \lambda\rho u = 0, \quad I^{(n_j)}(i, k_s), \quad u = 0 \quad \text{on } \partial I^{(n_j)}(i, k_s).$$

Then (3.1) implies $\lambda_1(I^{(n_j)}(i, k_1)) < \lambda_1(I^{(n_j)}(i, k_2))$, which is absurd, since both of these first eigenvalues are equal to λ_{n_j}, the eigenvalue of (1.2) corresponding to φ_{n_j}. This completes the proof of Theorem 3.1. □

If we know that the density function of (1.2) is radial: $\rho(x, y) = \rho(r)$, $r = (x^2 + y^2)^{1/2}$, then (1.2) has infinitely many eigenfunctions which are radial, and hence, circular. These eigenfunctions are eigenfunctions of the following eigenvalue problem:

$$u''(r) + \frac{u'(r)}{r} + \mu\rho(r)u(r) = 0, \qquad u(r_0) = u(r_1) = 0. \tag{3.2}$$

Let $r = e^t, \alpha = \ln r_0, \beta = \ln r_1, v(t) = u(r)$. Then (3.2) is transformed into the following eigenvalue problem:

$$v''(t) + \mu P(t)v(t) = 0, \qquad v(\alpha) = v(\beta) = 0, \tag{3.3}$$

where $P(t) = e^{2t}\rho(e^t)$. Let $v_n(t)$ be the nth eigenfunction, μ_n be the nth eigenvalue of (3.3). We will denote $x_1^{(n)} < x_2^{(n)} < \cdots < x_{n-1}^{(n)}$, the nodal points of v_n in the open interval (α, β), and we denote $x_0^{(n)} = \alpha$, $x_n^{(n)} = \beta$, $I_{n,j} = [x_{j-1}^{(n)}, x_j^{(n)}]$. Then by the variational formula for μ_n, and the monotonicity principle for the comparison of the eigenvalues, we have

$$\frac{\pi^2}{[\max(P, I_{n,j})|I_{n,j}|^2]} \leqq \mu_n \leqq \frac{\pi^2}{[\min(P, I_{n,j})|I_{n,j}|^2]}, \tag{3.4}$$

where $\max(P,I)$ (respectively, $\min(P,I)$) denotes the maximum (respectively, the minimum) of P on the interval I, and $|I|$ denotes the length of the interval I. The following lemma was proved in [5].

LEMMA 3.2. *For $\delta > 0$, there exists $n_0(\delta)$ such that $|I_{n,k}| < \delta, \quad k = 1,2,\dots,n$ for all $n \geqq n_0$.*

Using Lemma 3.2 we can prove the following theorem.

THEOREM 3.3. *Suppose that R is an annulus as above, and we know the density function ρ in* (1.2) *is radial. If we can "hear" infinitely many circular eigenvalues and we can "see" the circular nodal sets of the corresponding eigenfunctions, then we can determine the density function ρ.*

Proof. By assumption we know infinitely many eigenvalues μ_{n_k} of (3.3) and we know the corresponding nodal points $\alpha < x_1^{(n_k)} < \cdots < x_{n_k-1}^{(n_k)} < \beta$ of v_{n_k}. Then by (3.4), Lemma 3.2, and the fact that $P(t)$ is uniformly continuous on $[\alpha,\beta]$, for $x \in [\alpha,\beta]$, if $I_{n_k,j(x)}$ is the closure of the nodal domain of v_{n_k} which contains x in its interior or has x as its right end point, we then have

$$P(x) = \pi^2 \lim_{k\to\infty} \frac{1}{\mu_{n_k}|I_{n_k,j(x)}|^2}.$$

Hence $P(x)$ is determined, so is $\rho(r)$. □

We note that recently the problem of determining the coefficients of a Sturm–Liouville equation from the nodal sets of the eigenfunctions has attracted some attention (see [2], [3], and [5]). In particular, for the string equation with Neumann boundary condition, $y'' + \lambda\rho(x)y = 0$, $y'(0) = y'(L) = 0$, if the density function $\rho(x)$ has integrable second derivative, then it was shown by Hald and McLaughlin [2] that $\rho(x)$ can be determined uniquely up to a multiplicative constant by a dense set of nodal points of eigenfunctions. Their idea is to transform the string equation via the Liouville transform into an equation of the form $z''+(\lambda-q)z = 0$. Then they use the asymptotic behaviour of the nodal sets of the eigenfunctions of the latter to achieve their goal. Besides, Hald and McLaughlin also developed some interesting numerical methods for the inverse nodal problems. Our approach to the inverse problem of the eigenvalue problem (3.3) is different from theirs. We look at the nodal sets of the eigenfunctions of (3.3) directly, and we need eigenvalues.

Combining Theorems 3.1 and 3.3 we see that the appearance of infinitely many circular-nested eigenvalues such that the meshes of the nodal domains shrink to zero implies the density function ρ is radial. Then the appearance of circular nodal sets determines the density function ρ.

The following example shows that only the appearance of infinitely many circular nodal sets does not guarantee that the density function is radial. That is, to determine a radial density function, we need the appearance of both of circular and circular-nested eigenvalues and nodal sets.

Example. For (x,y), which lies on the circle $x^2 + y^2 = r^2$, $r > 0$, define $z = x + iy$, here $i = (-1)^{1/2}$, and define

$$w = w(z) = \frac{z + z^{-1}}{2} = u + iv, \qquad u = \operatorname{Re} w, \quad v = \operatorname{Im} w. \tag{3.5}$$

Then (u,v) lies on the ellipse E_r defined as follows:

$$E_r = \left\{(u,v) : \frac{u_2}{[(r+1/r)/2]^2} + \frac{v_2}{[(r-1/r)/2]^2} = 1\right\}. \tag{3.6}$$

Note that for $r > 0$, the foci of E_r are $(1,0)$, $(-1,0)$, i.e., E_r are confocal. Conversely, if E is an ellipse with foci $(\pm 1,0)$, then E is of the form E_r. For $1 < a < b$, via the map $w(z)$, the annulus $R(a,b) = \{a^2 < x^2 + y^2 < b\}$ is transformed into the doubly connected domain $E(a,b)$ bounded by two confocal ellipses, E_a and E_b. Via $w(z)$ eigenvalue problem

$$\Delta\varphi + \lambda\varphi = 0 \quad \text{in } E(a,b), \qquad \varphi = 0 \quad \text{on } \partial E(a,b), \tag{3.7}$$

is transformed into the following nonhomogeneous annular membrane problem:

$$\Delta\psi + \lambda\left\{\frac{|1 - 1/z^2|^2}{4}\right\}\psi = 0 \quad \text{in } R(a,b), \qquad \psi = 0 \quad \text{on } \partial R(a,b). \tag{3.8}$$

Suppose ξ_0, ξ_1 are chosen so that

$$\cosh^2\xi_0 = \left[\frac{(a+1/a)}{2}\right]^2, \qquad \cosh^2\xi_1 = \left[\frac{(b+1/b)}{2}\right]^2. \tag{3.9}$$

For $(x,y) \in E(a,b)$ we introduce the *elliptic-coordinate* (ξ,η) by the following formulae: $x = \cosh\xi\cos\eta, y = \sinh\xi\sin\eta$, where $\xi_0 < \xi < \xi_1, 0 \leqq \eta < 2\pi$. Then the functions

$$\Psi(\xi,\eta;q) := \left\{\left[\frac{\mathrm{Ce}_0(\xi,q)}{\mathrm{Ce}_0(\xi_0,q)}\right] - \left[\frac{\mathrm{Fey}_0(\xi,q)}{\mathrm{Fey}_0(\xi_0,q)}\right]\right\}\mathrm{ce}_0(\eta,q), \tag{3.10}$$

where q is determined by the following transcendental equation:

$$\frac{\mathrm{Ce}_0(\xi_1,q)}{\mathrm{Ce}_0(\xi_0,q)} - \frac{\mathrm{Fey}_0(\xi_1,q)}{\mathrm{Fey}_0(\xi_0,q)} = 0 \tag{3.11}$$

are eigenfunctions of (3.7) in elliptic coordinates. ce_0, Ce_0, and Fey_0 are Mathieu functions and modified Mathieu functions of the second kind (for definition and relevant properties; see [1]). Since $\mathrm{ce}_0(\eta,q) > 0$ in the interior of the elliptic annulus for the q determined by (3.11), the nodal set of $\Psi(\xi,\eta;q)$ are those of ξ_* such that

$$\frac{\mathrm{Ce}_0(\xi_*,q)}{\mathrm{Ce}_0(\xi_0,q)} - \frac{\mathrm{Fey}_0(\xi_*,q)}{\mathrm{Fey}_0(\xi_0,q)} = 0.$$

These nodal curves $\xi = \xi_*$ correspond to concentric circles in Euclidean coordinates. Thus the nonhomogeneous annular membrane problem (3.8) has infinitely many circular eigenfunctions, but the density function of (3.8) is not radial.

Thus the appearance of infinitely many circular eigenvalues does not guarantee that the density function of a nonhomogeneous annular membrane is radial.

Acknowledgments. The author wishes to thank the referee of this paper for referring him to the interesting paper of Hald and McLaughlin.

REFERENCES

[1] A. Erdelyi, Ed., *Higher Transcendental Functions*, Vol. III, McGraw-Hill, New York, 1955.

[2] O. H. Hald and J. R. McLaughlin, *Solutions of inverse nodal problems*, Inverse Problems, 5 (1989), pp. 307–347.

[3] J. R. McLaughlin, *Inverse spectral theory using nodal points as data – a uniqueness result*, J. Differential Equations, 73 (1988), pp. 354–362.

[4] C.-L. Shen, *Remarks on the second eigenfunction of a symmetric simply connected plane region*, SIAM J. Math. Anal., 19 (1988), pp. 167–171.

[5] ———, *On the nodal sets of the eigenfunctions of the string equation*, SIAM J. Math. Anal., 19 (1988), pp. 1419–1424.

SIAM J. MATH. ANAL.
Vol. 24, No. 5, pp. 1283–1293, September 1993

009

ON THE ZEROS OF SOLUTIONS TO GINZBURG–LANDAU TYPE SYSTEMS*

PATRICIA BAUMAN†‡, NEIL N. CARLSON†, AND DANIEL PHILLIPS†§

Abstract. The authors consider minimizers of a nonlinear functional whose Euler–Lagrange equation includes the Ginzburg–Landau system. For a certain class of Dirichlet data, it is proved that a minimizer has exactly one zero which necessarily has winding number ±1. Moreover, the same result holds for solutions of the corresponding parabolic system at all sufficiently large, fixed values of time, under certain conditions on the initial and boundary values.

Their result on minimizers supports several theories from physics (concerning interacting bosons, for example). These theories predict that *stable* solutions with isolated zeros (called vortices) exist, and each zero of a stable solution has winding number ±1.

Key words. Ginzburg–Landau system, vortices, zeros

AMS subject classifications. 35J55, 35Q40

1. Introduction.

Let Ω be a bounded simply connected domain in $\mathbb{R}^2$, and let $\mathcal{U}(X) : \Omega \to \mathbb{R}^2$ be a solution to the diagonal system

$$0 = \Delta \mathcal{U} - G'(|\mathcal{U}|^2) \cdot \mathcal{U}. \tag{1.1}$$

If $G'(|\mathcal{U}|^2) = (|\mathcal{U}|^2 - 1)$, this is the Ginzburg–Landau system used in both the theory for interacting bosons [Gr] and the theory of super-conductivity [G–L]. Here the squared modulus $|\mathcal{U}|^2$ plays the role of a density. Denote by $\Gamma(\mathcal{U})$ the set of zeros of $\mathcal{U}$, namely, $\Gamma(\mathcal{U}) = \{X \in \Omega : \mathcal{U}(X) = 0\}$. The above theories predict the existence of solutions for which $\Gamma(\mathcal{U})$ consists of isolated points called *vortices.* Moreover, it is expected that each vortex of a stable solution has winding number ±1.

In this paper we consider the variational problem associated with (1.1), subject to a particular class of Dirichlet data. We prove that an energy-minimizing solution has exactly one zero that necessarily has winding number ±1. This result then supports the theories described above.

To state our results more precisely, assume that Ω is of class $C^{2,\mu}$ for μ fixed with $0 < \mu < 1$. Let $Y(s)$ for $0 \leq s < L$ be a one-to-one parametrization of $\partial\Omega$ with respect to arclength. Consider Dirichlet data, ψ_0, in $C^{2,\mu}(\overline{\Omega}; \mathbb{R}^2)$. Expressing the image, $\psi_0(Y(s))$, of $\partial\Omega$ in polar coordinates, we have

$$\psi_0(Y(s)) = (r(s) \cdot \cos\theta(s), r(s) \cdot \sin\theta(s)).$$

We assume that ψ_0 satisfies

$$r(s) > 0, \theta'(s) \neq 0 \quad \text{for } 0 \leq s < L \quad \text{and} \quad |\theta(L) - \theta(0)| = 2\pi. \tag{1.2}$$

*Received by the editors April 13, 1992; accepted for publication November 16, 1992.

†Department of Mathematics, Purdue University, West Lafayette, Indiana 47907.

‡This author's research was partially supported by National Science Foundation grant DMS-9112884.

§This author's research was partially supported by National Science Foundation grant DMS-9112471.

Thus the image, $\psi_0(\partial\Omega)$, bounds a starlike region with respect to the origin in $\mathbb{R}^2$ such that $\psi_0(Y(s))$ crosses each ray, $\theta = \theta_0$, exactly once as s increases from zero to L. We assume that the nonlinearity, G, satisfies

$$G \in C^2([0,\infty)), \quad G \geq 0, \quad \text{and} \quad \overline{\lim_{\tau\to\infty}} \frac{|G'(\tau)|}{\tau^\lambda} < \infty \quad \text{for some } \lambda > 0. \tag{1.3}$$

Consider the variational problem of minimizing the functional

$$J(\psi) = \int_\Omega (|\nabla\psi|^2 + G(|\psi|^2))dX \tag{1.4}$$

in the class $\mathcal{M} \equiv \{\psi \in W^{1,2}(\Omega;\mathbb{R}^2) : \psi = \psi_0 \text{ on } \partial\Omega\}$. Our main result is the following.

THEOREM. *If $\mathcal{U}$ is a minimizer of $J(\cdot)$ in $\mathcal{M}$, then $\Gamma(\mathcal{U})$ consists of one point.* (*See Theorem* 2.3.)

In general, the most one can say about $\Gamma(\psi)$ for a given continuous vector field ψ with $\psi = \psi_0$ on $\partial\Omega$ is that its Brouwer degree satisfies $d(\psi,\Omega,0) = \pm 1$. Thus, our result states that a minimizer has the simplest structure of this type possible.

An interesting case is with $\Omega = B_1(0)$ and $\psi_0(X) = X$. It is known that there exists a separable solution of (1.1) with boundary values, ψ_0, namely,

$$\mathcal{U}_R(X) = R(|X|)(X/|X|),$$

with $R(0) = 0$, $R(1) = 1$, and $0 < R(|X|) < 1$ for $|X|$ in $(0,1)$. See [Ab]. It is not known, however, if $\mathcal{U}_R$ is a minimizer of $J(\cdot)$ in all of $\mathcal{M}$. Thus the fact that a minimizer has one zero is new even in this case.

The stationary result can be used to investigate the pattern of the zero set as it evolves with time for the nonlinear heat equation studied in [N]. Consider the system

$$\begin{aligned} \mathcal{U}_t &= \Delta\mathcal{U} - G'(|\mathcal{U}|^2)\mathcal{U} && \text{for } X \in \Omega, \quad t > 0, \\ \mathcal{U}(X,t) &= \psi_0(X) && \text{for } X \in \partial\Omega, \quad t > 0, \\ \mathcal{U}(X,0) &= \mathcal{U}_0(X) && \text{for } X \in \Omega. \end{aligned} \tag{1.5}$$

It is known that if $\mathcal{U}_0 \in \mathcal{M}$, then there is a unique solution $\mathcal{U} = \mathcal{U}(t)$, which is classical for $t > 0$. Setting $m = \inf_{\psi\in\mathcal{M}} J(\psi)$ we have the following.

THEOREM. *There is a constant $\delta > 0$ such that if $J(\mathcal{U}_0) \leq m + \delta$, then $\Gamma(\mathcal{U}(t)) = \{X(t)\}$, where $X(t)$ is a C^1 curve in Ω for all t sufficiently large.* (*See Theorem* 3.3.)

If $G(\tau)$ is real analytic, with Ω and ψ_0 sufficiently smooth, then more can be said. A result of Simon states that there is a constant $\delta_1 > 0$ such that if $J(\mathcal{U}_0) \leq m + \delta_1$, then $\mathcal{U}(t) \to \mathcal{U}^\infty$ in $C^2(\overline{\Omega};\mathbb{R}^2)$ as $t \to \infty$, where $\mathcal{U}^\infty$ is a minimizer. We use this to prove the following.

THEOREM. *$X(t) \to X^\infty$ as $t \to \infty$, where $\{X^\infty\} = \Gamma(\mathcal{U}^\infty)$.* (*See Theorem* 3.6.)

Thus in this setting there is large-time dynamic stability for the pattern of $\Gamma(\mathcal{U}(t))$ near the minimum energy level.

Notation. When the context is clear we use the following abbreviated notation:

$$\begin{gathered} W^{1,2}(\Omega;\mathbb{R}^2) = W^{1,2}(\Omega), \\ C^{2,\mu}(\overline{\Omega};\mathbb{R}^2) = C^{2,\mu}(\overline{\Omega}), \\ \|\mathcal{U}\|_{C^{2,\mu}(\overline{\Omega};\mathbb{R}^2)} = \|\mathcal{U}\|_{C^{2,\mu}}. \end{gathered}$$

2. Minimizers and their zeros. We first summarize the existence and regularity results related to the variational problem.

LEMMA 2.1. *There is at least one minimizer for $J(\cdot)$ in $\mathcal{M}$. A minimizer, $\mathcal{U} = (u,v)$, is a solution of*

$$\begin{aligned} \Delta u - G'(|\mathcal{U}|^2)u &= 0, \\ \Delta v - G'(|\mathcal{U}|^2)v &= 0 \quad \text{in } \Omega, \\ (u,v) &= \psi_0 \quad \text{on } \partial\Omega. \end{aligned} \tag{2.1}$$

Any weak solution, $\widetilde{\mathcal{U}}$, of this system in $W^{1,2}(\Omega)$ is of class $C^{2,\mu}(\overline{\Omega})$ and

$$\|\widetilde{\mathcal{U}}\|_{C^{2,\mu}(\overline{\Omega})} \le C\left(\|\widetilde{\mathcal{U}}\|_{W^{1,2}(\Omega)}, \|\psi_0\|_{C^{2,\mu}(\overline{\Omega})}\right). \tag{2.2}$$

Proof. Since G satisfies the conditions in (1.3), the general theory of variational problems implies the existence of a minimizer, $\mathcal{U}$, in $\mathcal{M}$. See [G, Chap. I]. Since $\Omega \subset \mathbb{R}^2$ and $\mathcal{U} \in W^{1,2}(\Omega)$, it follows from the Sobolev imbedding theorem that $\mathcal{U} \in L^p(\Omega)$ for any $p < \infty$. From this and (1.3), $J(\mathcal{U}+\varepsilon\Phi)$ is differentiable in ε for any $\Phi \in C_c^1(\Omega;\mathbb{R}^2)$. Thus $\mathcal{U}$ is a weak solution of (2.1).

Now let $\widetilde{\mathcal{U}} = (\widetilde{u},\widetilde{v})$ be any weak solution of (2.1) in $W^{1,2}(\Omega)$. By (1.3), (2.1), and the Sobolev imbedding theorem $\Delta\widetilde{u}$ and $\Delta\widetilde{v}$ are in $L^q(\Omega)$ for any $q < \infty$. Hence by elliptic estimates,

$$\begin{aligned} \|\widetilde{\mathcal{U}}\|_{C^1(\overline{\Omega})} &\le C(\|\widetilde{\mathcal{U}}\|_{L^q(\Omega)}, \|\psi_0\|_{C^{2,\mu}(\overline{\Omega})}) \\ &\le \widetilde{C}(\|\widetilde{\mathcal{U}}\|_{W^{1,2}(\Omega)}, \|\psi_0\|_{C^{2,\mu}(\overline{\Omega})}) \end{aligned}$$

for all q sufficiently large. Thus $\Delta\widetilde{u}$, $\Delta\widetilde{v} \in C^{\mu}(\overline{\Omega})$, and (2.2) follows. □

For $\alpha \in \mathbb{R}$ and $\mathcal{U} = (u,v)$, a minimizer for $J(\cdot)$ in $\mathcal{M}$, set

$$w_\alpha(X) = -u(X)\cdot\sin\alpha + v(X)\cdot\cos\alpha.$$

Define the *nodal set* of w_α by

$$N_\alpha \equiv \{X \in \overline{\Omega} : w_\alpha(X) = 0\}.$$

Note that for any pair α_1, α_2 with $0 \le \alpha_1 < \alpha_2 < \pi$, the set of zeros of $\mathcal{U}$ is just $\Gamma(\mathcal{U}) = N_{\alpha_1} \cap N_{\alpha_2}$. Thus as α varies, $\Gamma(\mathcal{U})$ is the subset of N_α that remains fixed.

We shall prove (in Lemma 2.2) that for each α, N_α is a smooth imbedded curve which enters and exits $\overline{\Omega}$ at distinct points of $\partial\Omega$. Following this, we prove that the curves, N_α, have exactly one point in common.

LEMMA 2.2. *For each α, N_α is a C^1 imbedded curve in $\overline{\Omega}$.*

Proof. First consider $N_\alpha \cap \partial\Omega$. From (1.2) we have $N_\alpha \cap \partial\Omega = \{P_1, P_2\}$. Using the notation from (1.2) if $Y(s_1) = P_1$ and $Y(s_2) = P_2$, we can assume without loss of generality that $\theta(s_1) = \alpha + \pi$ and $\theta(s_2) = \alpha$. Now

$$w_\alpha(Y(s)) = r(s)\cdot[-\cos\theta(s)\cdot\sin\alpha + \sin\theta(s)\cdot\cos\alpha].$$

Hence

$$\left.\frac{\partial}{\partial s}w_\alpha(Y(s))\right|_{s_1} = -r(s_1)\cdot\theta'(s_1) \ne 0$$

and

$$\frac{\partial}{\partial s} w_\alpha(Y(s))\bigg|_{s_2} = r(s_2)\cdot\theta'(s_2) \neq 0.$$

Thus there are neighborhoods $\mathcal{O}_1$ and $\mathcal{O}_2$ of P_1 and P_2, respectively, so that $N_\alpha \cap \mathcal{O}_1$ and $N_\alpha \cap \mathcal{O}_2$ are C^1 curves intersecting $\partial\Omega$ at P_1 and P_2.

Next we examine N_α in Ω. Note that w_α is a $C^{2,\mu}$ solution of

$$\Delta w_\alpha - G'(|\mathcal{U}|^2)w_\alpha = 0 \quad \text{in } \Omega,$$

where $G'(|\mathcal{U}|^2)$ is continuous. By Hartman and Wintner's classical results (see [H-W, Thms. 1–2 and Cor. 1]), the set

$$K_\alpha \equiv \{X \in \Omega : w_\alpha(X) = 0 \text{ and } \nabla w_\alpha(X) = 0\}$$

is locally finite. Our previous analysis near $\partial\Omega$ then implies that K_α is either empty or is a finite subset of Ω. It follows from [H-W] and our analysis near $\partial\Omega$ that N_α consists of a finite number of C^1 arcs along which $\nabla w_\alpha \neq 0$ except at their endpoints in Ω; moreover, the arcs may intersect only at these (interior) endpoints. Exactly two endpoints of these arcs are in $\partial\Omega$, and the rest make up K_α.

Finally, we note that at least four distinct arcs in N_α meet at each point in K_α. This follows from Hartman and Wintner's analysis of w_α near X_0 in K_α: Indeed, they show that for some integer n there is a homogeneous harmonic polynomial, H_n, of order n so that

$$w_\alpha(X) - H_n(X - X_0) = o(|X - X_0|^n) \quad \text{and}$$
$$\nabla w_\alpha(X) - \nabla H_n(X - X_0) = o(|X - X_0|^{n-1}).$$

(See (5) and (5′) of §1 in [H-W].) It follows that the nodal set of w_α has the same structure near X_0 as that of the harmonic function, $H_n(X - X_0)$.

We claim that $K_\alpha = \emptyset$. If not, fix X_0 in K_α. Let C_1, C_2, and C_3 be three distinct maximal (piecewise C^1) curves in N_α with endpoints at X_0. If *none* of these curves returns to X_0 or self-intersects or intersects one of the other two curves, then all three curves must reach $\partial\Omega$. But this is impossible by our analysis of N_α near $\partial\Omega$. Thus there exists a nontrivial subdomain $D_\alpha \subset \Omega$ with $\partial D_\alpha \subset N_\alpha$. Using a rotated basis for the image, we can represent $\mathcal{U}$ by $\mathcal{U}(X) = (w_\alpha(X), w_{\alpha+\frac{\pi}{2}}(X))$. Let

$$\widetilde{\mathcal{U}} = \begin{cases} \mathcal{U} & \text{in } \Omega - D_\alpha, \\ (-w_\alpha, w_{\alpha+\frac{\pi}{2}}) & \text{in } \overline{D}_\alpha. \end{cases}$$

Then $\widetilde{\mathcal{U}} \in \mathcal{M}$ and $J(\mathcal{U}) = J(\widetilde{\mathcal{U}})$. Thus $\widetilde{\mathcal{U}}$ is also a minimizer and (by Lemma 2.1) $\widetilde{\mathcal{U}} \in C^{2,\mu}(\overline{\Omega})$. But this is possible only if $\nabla w_\alpha \equiv 0$ on ∂D_α, which contradicts the fact that K_α is a finite set. This proves that $K_\alpha = \emptyset$. The same argument implies N_α does not contain a closed loop. Hence N_α is connected.

It follows that $\nabla w_\alpha \neq 0$ on N_α and N_α can be parametrized by arclength as a smooth imbedded curve in $\overline{\Omega}$ with endpoints $\{P_1, P_2\}$. □

The set, $\Gamma(\mathcal{U})$, is nonempty since $d(\mathcal{U}, \Omega, 0) = \pm 1$. In the following theorem we construct a homotopy from N_0 to N_π, passing through each N_α for $0 \leq \alpha \leq \pi$. By analyzing the points contained in all the N_α we prove that $\Gamma(\mathcal{U})$ has just one element.

THEOREM 2.3. *Let $\mathcal{U}$ be a minimizer for $J(\cdot)$ in $\mathcal{M}$ where ψ_0 satisfies* (1.2). *Then $\Gamma(\mathcal{U}) = \{X_0\}$ for some X_0 in Ω.*

Proof. Let $\{P_1, P_2\} = N_0 \cap \partial\Omega$ and assume without loss of generality that $\mathcal{U}(P_1) = (u(P_1), 0)$ with $u(P_1) < 0$ and $\mathcal{U}(P_2) = (u(P_2), 0)$ with $u(P_2) > 0$. The points, P_1 and P_2, partition $\partial\Omega$ into two arcs, $(\partial\Omega)^- \equiv \{X \in \partial\Omega : v(X) \le 0\}$ and $(\partial\Omega)^+ \equiv \{X \in \partial\Omega : v(X) \ge 0\}$. Starting at P_1 and moving along N_0, let X_0 be the first point reached in $\Gamma(\mathcal{U})$. Since $X_0 \in N_\alpha$ for all α, it follows from Lemma 2.2 that each N_α can be parametrized by arclength, $X = X(\tau, \alpha)$, such that

$$a(\alpha) < \tau < b(\alpha), \quad a(\alpha) < 0, \quad X(a(\alpha), \alpha) \in (\partial\Omega)^-, \quad b(\alpha) > 0, \quad X(b(\alpha), \alpha) \in (\partial\Omega)^+,$$
$$\left|\frac{\partial X}{\partial \tau}\right| = 1 \quad \text{and} \quad X(0, \alpha) = X_0.$$

Set $\mathcal{D} = \{(\tau, \alpha) : a(\alpha) \le \tau \le b(\alpha), 0 \le \alpha \le \pi\}$. We split the proof into three parts.

Part 1. We first show that $X \in C^1(\mathcal{D})$, and that $a(\alpha)$ and $b(\alpha)$ are in $C^1([0, \pi])$. Without loss of generality, assume that

$$\frac{\partial X}{\partial \tau}(0,0) = \left(\frac{\partial w_0}{\partial x_2}(X_0), -\frac{\partial w_0}{\partial x_1}(X_0)\right) \Big/ |\nabla w_0(X_0)|.$$

By Lemma 2.2 and the equicontinuity of $\{\nabla w_\alpha(X) : 0 \le \alpha \le \pi\}$, we have

$$\inf\{|\nabla w_\alpha(X)| : X \in N_\alpha \text{ and } 0 \le \alpha \le \pi\} > 0.$$

Thus from the theory of ordinary differential equations, it follows that $X(\tau, \alpha)$ solves the initial value problem

$$\begin{aligned} \frac{\partial x_1}{\partial \tau} &= \frac{\partial w_\alpha}{\partial x_2}(x_1, x_2) \Big/ |\nabla w_\alpha(x_1, x_2)|, \\ \frac{\partial x_2}{\partial \tau} &= -\frac{\partial w_\alpha}{\partial x_1}(x_1, x_2) \Big/ |\nabla w_\alpha(x_1, x_2)|, \\ (x_1(0,\alpha)&, x_2(0, \alpha)) = X_0. \end{aligned}$$

Since $w_\alpha(X)$ is a C^2 function of α and X it follows that $X(\tau, \alpha) \in C^1(\mathcal{D})$.

To show that $a(\alpha)$ is C^1, let $Q_1 \in \partial\Omega$ with $Q_1 = X(a(\alpha_1), \alpha_1)$. Choose Φ so that the equation, $\Phi(X) = 0$, defines $\partial\Omega$ in a neighborhood of Q_1 and $\nabla\Phi(Q_1) \ne 0$. In the proof of Lemma 2.2 we showed that the tangential component of $\nabla w_\alpha(Q_1)$ does not vanish. It follows that $\nabla\Phi(Q_1) \cdot (\partial X/\partial\tau)(a(\alpha_1), \alpha_1) \ne 0$. Thus the equation $\Phi(X(\tau, \alpha)) = 0$ implicitly determines $\tau = a(\alpha)$ as a C^1 function of α near α_1. Hence $a(\alpha) \in C^1([0, \pi])$. In the same manner, we have $b(\alpha) \in C^1([0, \pi])$. Thus $X(\tau, \alpha)$ is a C^1 homotopy defined on the piecewise smooth domain $\mathcal{D}$.

Part 2. By construction, $X(\tau, 0) \notin \Gamma(\mathcal{U})$ for $\tau < 0$. Here we show that for all α in $[0, \pi]$ and $\tau < 0$, $X(\tau, \alpha) \notin \Gamma(\mathcal{U})$. In particular, $X(\tau, \pi) \notin \Gamma(\mathcal{U})$ for $\tau < 0$.

Set
$$\overline{\alpha} = \sup\{\alpha : X(\tau, \beta) \notin \Gamma(\mathcal{U}) \text{ for } \tau < 0 \text{ and } \beta \le \alpha\}.$$

There are three possibilities. The first is that there exists $\overline{\tau} < 0$ such that $X(\overline{\tau}, \overline{\alpha}) \in \Gamma(\mathcal{U})$. Since $\Gamma(\mathcal{U}) \subset N_\alpha$ for all α, we have $X(\overline{\tau}, \overline{\alpha}) \in N_\alpha$ for all $\alpha < \overline{\alpha}$. By construction, for each $\alpha < \overline{\alpha}$ there exists $\tau(\alpha) \in (0, b(\alpha))$ with $X(\tau(\alpha), \alpha) = X(\overline{\tau}, \overline{\alpha})$. By continuity this implies that for some $\tau^* \in [0, b(\alpha)]$, we have $X(\tau^*, \overline{\alpha}) = X(\overline{\tau}, \overline{\alpha})$. But $N_{\overline{\alpha}}$ does not self-intersect (by the proof of Theorem 2.2). Hence this case is impossible.

The second possibility is that $\overline{\alpha} < \pi$ and $X(\tau, \overline{\alpha}) \notin \Gamma(\mathcal{U})$ for all $\tau < 0$. In this case, there must be a sequence, $\{\tau_n, \alpha_n\}$, with $\alpha_n \downarrow \overline{\alpha}$, $\tau_n \uparrow 0$, and $X(\tau_n, \alpha_n)$ in $\Gamma(\mathcal{U})$. By continuity, $X(\tau_n, \alpha_n) \to X(0, \overline{\alpha}) = X_0$. Moreover, since $\Gamma(\mathcal{U}) \subset N_{\overline{\alpha}}$, there is a sequence $\{t_n\}$ with $t_n > 0$ so that $X(t_n, \overline{\alpha}) = X(\tau_n, \alpha_n)$. Necessarily, $t_n \downarrow 0$. Consider the vectors

$$\frac{X(\tau_n, \alpha_n) - X(0, \alpha_n)}{\tau_n}$$

and

$$\frac{X(t_n, \overline{\alpha}) - X(0, \overline{\alpha})}{t_n}.$$

Since $X \in C^1(\mathcal{D})$ both sequences tend to $(\partial X/\partial \tau)(0, \overline{\alpha})$ as $n \to \infty$. But for each n the angle between the two vectors is π. Hence $(\partial X/\partial \tau)(0, \overline{\alpha}) = 0$. This contradicts the fact that $|\partial X/\partial \tau| = 1$.

The only remaining possibility is that $\overline{\alpha} = \pi$ and $X(\tau, \pi) \notin \Gamma(\mathcal{U})$ for $\tau < 0$.

Part 3. Now $X(\tau, 0)$ and $X(\tau, \pi)$ are each parametrizations of N_0 in opposite directions. Since X_0 is the first point in each direction along N_0 that is in $\Gamma(\mathcal{U})$, we conclude that $\Gamma(\mathcal{U}) = \{X_0\}$. □

Remark. The idea of using a homotopy of nodal curves based on the result of Lemma 2.2 was motivated by an argument of Payne in [P] where he shows that in a convex domain in $\mathbb{R}^2$, an eigenfunction for the first eigenvalue of the Laplacian has one critical point.

COROLLARY 2.4. *Let $\mathcal{U}$ be a minimizer for $J(\cdot)$ in $\mathcal{M}$. The curves $\{u = 0\}$ and $\{v = 0\}$ cross only at $\Gamma(\mathcal{U}) = \{X_0\}$ in Ω. In fact, there is a constant $c > 0$ independent of $\mathcal{U}$ such that $|\det \nabla\mathcal{U}(X_0)| \geq c$.*

Proof. If there is no such lower bound, then there exists a sequence of minimizers $\{\mathcal{U}^n\}$ with $\mathcal{U}^n(X_0^n) = 0$, and $\det \nabla\mathcal{U}^n(X_0^n) \to 0$. From the uniform bound (2.2) for minimizers, it follows that there exists a minimizer, $\widetilde{\mathcal{U}}$, and a subsequence, $\{\mathcal{U}^j\}$, such that $\mathcal{U}^j \to \widetilde{\mathcal{U}}$ in $C^2(\overline{\Omega})$, $X_0^j \to \widetilde{X}_0 \in \Omega$, $\widetilde{\mathcal{U}}(\widetilde{X}_0) = 0$, and $\det \nabla\widetilde{\mathcal{U}}(\widetilde{X}_0) = 0$. Let α be such that $[-\sin\alpha, \cos\alpha] \cdot \nabla\widetilde{\mathcal{U}}(X_0) = 0$. Consider the function $\widetilde{w}_\alpha = -\widetilde{u}\sin\alpha + \widetilde{v}\cos\alpha$. We have $\nabla\widetilde{w}_\alpha(\widetilde{X}_0) = 0$ and $\widetilde{w}_\alpha(\widetilde{X}_0) = 0$. This is impossible by the proof of Lemma 2.2. □

3. The time dependent problem. In this section we consider the set of zeros of solutions to the parabolic problem,

$$\text{(3.1)} \qquad \begin{aligned} \mathcal{U}_t &= \Delta\mathcal{U} - G'(|\mathcal{U}|^2)\mathcal{U} && \text{for } X \in \Omega, \quad t > 0, \\ \mathcal{U}(X,t) &= \psi_0(X) && \text{for } X \in \partial\Omega, \quad t > 0, \\ \mathcal{U}(X,0) &= \mathcal{U}_0(X) && \text{for } X \in \overline{\Omega}, \end{aligned}$$

where $\psi_0 \in C^{2,\mu}(\overline{\Omega}; \mathbb{R}^2)$ and $\mathcal{U}_0 \in \mathcal{M} \equiv \{\psi \in W^{1,2}(\Omega; \mathbb{R}^2) : \psi = \psi_0 \text{ on } \partial\Omega\}$. We prove that for ψ_0 as before and for certain initial data, $\mathcal{U}_0$, the zero set of $\mathcal{U}$, namely, $\Gamma(\mathcal{U}(t)) = \{X \in \Omega : \mathcal{U}(X,t) = 0\}$, consists of exactly one point for all t sufficiently large. (See Theorem 3.3.)

We begin by stating existence, uniqueness, and regularity results for this problem.

THEOREM 3.1. *Assume $\psi_0 \in C^{2,\mu}(\overline{\Omega})$ and $\mathcal{U}_0 \in \mathcal{M}$. There exists a unique solution, $\mathcal{U}(t) \equiv \mathcal{U}(X,t)$, of* (3.1) *in $C([0,\infty); W^{1,2}(\Omega;\mathbb{R}^2))$. The solution is classical for $t > 0$ and*

$$\|\mathcal{U}(t)\|_{C^{2,\mu}(\overline{\Omega})} \le M_1 \quad \text{for } t \ge 1, \tag{3.2}$$

where M_1 depends only on Ω, μ, G, $J(\mathcal{U}_0)$, and $\|\psi_0\|_{C^{2,\mu}(\overline{\Omega})}$. Moreover, if $k \ge 3$, $G \in C^{k+1}([0,\infty))$, Ω is of class $C^{k,\mu}$, and $\psi_0 \in C^{k,\mu}(\overline{\Omega};\mathbb{R}^2)$, then $\mathcal{U}(t) \in C^{k,\mu}(\overline{\Omega};\mathbb{R}^2)$ for each $t > 0$ and

$$\|\mathcal{U}(t)\|_{C^{k,\mu}(\overline{\Omega})} \le M_2 \quad \text{for } t \ge 1. \tag{3.3}$$

The constant, M_2, depends only on Ω, μ, k, G, $J(\mathcal{U}_0)$, and $\|\psi_0\|_{C^{k,\mu}(\overline{\Omega})}$.

Proof. Let $\mathcal{V}$ be the unique solution of $\Delta\mathcal{V} = 0$ in Ω, $\mathcal{V} = \psi_0$ on $\partial\Omega$. Set $\widetilde{\mathcal{U}}(X,t) = \mathcal{U}(X,t) - \mathcal{V}(X)$. By elliptic estimates, $\mathcal{V} \in C^{2,\mu}(\overline{\Omega})$ and $\|\mathcal{V}\|_{C^{2,\mu}(\overline{\Omega})} \le M_1$ for some constant M_1 as above. Moreover, if Ω is $C^{k,\mu}$ and $\psi_0 \in C^{k,\mu}(\overline{\Omega})$ for $k \ge 3$, then $\|\mathcal{V}\|_{C^{k,\mu}(\overline{\Omega})} \le M_2$ for some constant M_2 as above. Thus the theorem follows if we prove that there is a unique solution, $\widetilde{\mathcal{U}}(t) \equiv \widetilde{\mathcal{U}}(X,t)$, in $C([0,\infty); W^{1,2}(\Omega))$ of the boundary value problem

$$\begin{aligned}
\widetilde{\mathcal{U}}_t &= \Delta\widetilde{\mathcal{U}} - G'(|\widetilde{\mathcal{U}} + \mathcal{V}|^2)\cdot(\widetilde{\mathcal{U}} + \mathcal{V}) && \\
&\equiv \Delta\widetilde{\mathcal{U}} + F(\widetilde{\mathcal{U}}), && \text{for } X \in \Omega, \quad t > 0,\\
\widetilde{\mathcal{U}}(X,t) &= 0, && \text{for } X \in \partial\Omega, \quad t > 0,\\
\widetilde{\mathcal{U}}(X,0) &= \mathcal{U}_0(X) - \mathcal{V}(X) && \text{for } X \in \overline{\Omega},
\end{aligned} \tag{3.4}$$

and that $\widetilde{\mathcal{U}}$ is classical and satisfies (3.2) and (3.3).

To prove existence and uniqueness of solutions to (3.4), we apply semigroup theory as in [H]. Fix $p \ge 2$ and note that $F(\psi) \equiv -G'(|\psi + \mathcal{V}|^2)(\psi + \mathcal{V})$ is locally Lipschitz continuous from $W_0^{1,p}(\Omega) \equiv W_0^{1,p}(\Omega;\mathbb{R}^2)$ into $L^p(\Omega) \equiv L^p(\Omega;\mathbb{R}^2)$. Let $A : W_0^{1,p}(\Omega) \cap W^{2,p}(\Omega) \to L^p(\Omega)$ be defined by $A(\psi) = -\Delta\psi$. Then $A = A_p$ is a densely defined, positive, self-adjoint operator on $X \equiv L^p(\Omega)$ and A^{-1} is a compact operator. The operator $A^{1/2}$ is defined on $X^{1/2} \equiv D(A^{1/2}) = W_0^{1,p}(\Omega)$.

Consider the equations

$$\begin{aligned}
\widetilde{\mathcal{U}}_t &= -A\widetilde{\mathcal{U}} + F(\widetilde{\mathcal{U}}) \quad \text{in } L^p(\Omega) \quad \text{for } t > t_0 \ge 0,\\
\widetilde{\mathcal{U}}(t_0) &= \widetilde{\mathcal{U}}_0.
\end{aligned} \tag{3.5}$$

If $\widetilde{\mathcal{U}}_0 \in W_0^{1,p}(\Omega)$, then the existence-uniqueness result of Henry (see Theorem 3.3 of [H]) states that there exists $t_1 > t_0$ such that (3.5) has a unique solution on (t_0, t_1). Here, such a solution is defined as a mapping $\widetilde{\mathcal{U}}$, in $C([t_0,t_1); L^p(\Omega))$ such that $\widetilde{\mathcal{U}}(t) \in W_0^{1,p}(\Omega) \cap W^{2,p}(\Omega) \equiv D(A)$, $\partial\widetilde{\mathcal{U}}/\partial t$ exists, $t \to F(\widetilde{\mathcal{U}}(t))$ is locally Hölder continuous on (t_0, t_1),

$$\int_{t_0}^{t_0+\rho} \|F(\widetilde{\mathcal{U}}(t))\|_{L^p(\Omega)} dt < \infty \quad \text{for some } \rho > 0,$$

$\widetilde{\mathcal{U}}(t_0) = \widetilde{\mathcal{U}}_0$, and $\widetilde{\mathcal{U}}$ satisfies (3.5) in $L^p(\Omega)$ for $t_0 < t < t_1$.

Applying the above result with $p = 2$, $t_0 = 0$, and $\widetilde{\mathcal{U}}_0 = \mathcal{U}_0 - \mathcal{V}$, we obtain a unique solution of (3.5) in $C([0,t_1); L^2(\Omega))$ for some $t_1 > 0$.

We first prove that $t_1 = \infty$. Choose $t_2 \geq t_1$ such that $(0,t_2)$ is the maximal interval of existence-uniqueness of the solution in $L^2(\Omega)$. Fix $p > 2$, and let K_1 denote constants depending only on Ω, p, μ, G, $J(\mathcal{U}_0)$, and $\|\psi_0\|_{C^{2,\mu}(\overline{\Omega})}$. By standard energy estimates (see (3.11)),

$$J(\mathcal{U}(t)) \leq J(\mathcal{U}(0)) \quad \text{for all } t \text{ in } (0,t_2),$$

where $\mathcal{U} = \widetilde{\mathcal{U}} + \mathcal{V}$. Since $W_0^{1,2}(\Omega) \subset L^p(\Omega)$ for all $p \geq 2$, we have

$$\begin{aligned}(3.6)\qquad \|\widetilde{\mathcal{U}}(t)\|_{L^p(\Omega)} &\leq K_1 \cdot \|\widetilde{\mathcal{U}}(t)\|_{W_0^{1,2}(\Omega)}\\ &\leq K_1 \cdot J(\mathcal{U}(t)) + \|\mathcal{V}\|_{W_0^{1,2}(\Omega)}\\ &\leq K_1 \cdot (J(\mathcal{U}(0)) + 1) \leq K_1\end{aligned}$$

for all t in $(0,t_2)$. By (1.3) and the above inequality, we have

$$(3.7)\qquad \|F(\widetilde{\mathcal{U}}(t))\|_{L^p(\Omega)} = \|G'(|(\widetilde{\mathcal{U}} + \mathcal{V})(t)|^2) \cdot (\widetilde{\mathcal{U}} + \mathcal{V})(t)\|_{L^p(\Omega)} \leq K_1$$

for all t in $(0,t_2)$. Now $\widetilde{\mathcal{U}}(T) \in D(A) = W_0^{1,2}(\Omega) \cap W^{2,2}(\Omega) \subset W_0^{1,p}(\Omega)$ for any T in $(0,t_2)$. Thus by the existence-uniqueness result in [H],

$$(3.8)\qquad \widetilde{\mathcal{U}}(t) = e^{-A(t-T)}\widetilde{\mathcal{U}}(T) + \int_T^t e^{-A(t-s)}F(\widetilde{\mathcal{U}}(s))ds$$

in $W_0^{1,p}(\Omega)$ for all t in (T,T_1), where $T < T_1 < t_2$ and T_1 depends on Ω, p, F, T, and $\widetilde{\mathcal{U}}(T)$. By a result of Alikakos (see [Al, p. 285]), it follows from (3.6)–(3.8) that

$$(3.9)\qquad \|\widetilde{\mathcal{U}}(t)\|_{W^{1,p}(\Omega)} \leq K_1 \quad \text{for } T < t < T_1,$$

and we may take $T_1 = t_2 = t_1 = \infty$.

Finally, to show that $\widetilde{\mathcal{U}}$ is classical, we note that by (3.9) with $p = p_0 > 2$,

$$\|\widetilde{\mathcal{U}}(t)\|_{L^\infty(\Omega)} \leq K_1 \quad \text{for all } t > 0,$$

where K_1 is independent of t. It follows from parabolic estimates (see [L-S-U, Chap. 2, Lem. 3.3 and Chap. 4, Thms. 5.2 and 9.1]) that $\widetilde{\mathcal{U}}$ is classical and

$$\|\widetilde{\mathcal{U}}(t)\|_{C^{2,\mu}(\overline{\Omega})} \leq M_1 \quad \text{for } t \geq t_0 > 0,$$

where M_1 depends only on K_1 and t_0. Moreover, if $k \geq 2$, $G \in C^{k+1}([0,\infty))$, Ω is of class $C^{k,\mu}$, and $\psi_0 \in C^{k,\mu}(\overline{\Omega})$, then $\mathcal{U}(t) \in C^{k,\mu}(\overline{\Omega})$ for $t > 0$ and

$$\|\mathcal{U}(t)\|_{C^{k,\mu}(\overline{\Omega})} \leq M_2 \quad \text{for } t \geq t_0,$$

where M_2 depends only on t_0, k, Ω, μ, G, $J(\mathcal{U}_0)$, and $\|\psi\|_{C^{k,\mu}(\overline{\Omega})}$. □

DEFINITION. If $\mathcal{F} \subset C^2(\overline{\Omega}; \mathbb{R}^2)$ and $\varepsilon > 0$, set

$$\mathcal{F}_\varepsilon = \{\mathcal{V} \in C^2(\overline{\Omega}) : \|\mathcal{V} - \mathcal{U}\|_{C^2(\overline{\Omega})} < \varepsilon \text{ for some } \mathcal{U} \in \mathcal{F}\}.$$

Fix ψ_0 in $C^{2,\mu}(\overline{\Omega})$ such that ψ_0 satisfies (1.2). Let $\mathcal{S}$ be the set of all minimizers of J in $\mathcal{M}$. By Theorem 2.1, $\mathcal{S}$ is a bounded subset of $C^2(\overline{\Omega})$. We shall need the following lemma concerning $\mathcal{S}_\varepsilon$.

LEMMA 3.2. *There exists $\varepsilon > 0$ such that if $\mathcal{V} \in \mathcal{S}_\varepsilon$, then $\Gamma(\mathcal{V}) \equiv \{X \in \Omega : \mathcal{V}(X) = 0\}$ is a single point.*

Proof. From (2.2) and Corollary 2.4, there exist positive constants C_1 and C_2 (independent of $\mathcal{U}$) such that if $\mathcal{U} \in \mathcal{S}$ and $\{X_0\} = \Gamma(\mathcal{U})$, then

$$\|\mathcal{U}\|_{C^2(\overline{\Omega})} \leq C_1 \quad \text{and} \quad |\det \nabla\mathcal{U}(X_0)| \geq C_2.$$

It follows that if ε_1 is sufficiently small (depending on C_1 and C_2), $\mathcal{V} \in \mathcal{S}_{\varepsilon_1}$, and $\|\mathcal{V} - \mathcal{U}\|_{C^2(\overline{\Omega})} < \varepsilon_1$, then

$$\|\mathcal{V}\|_{C^2(\overline{\Omega})} \leq 2C_1 \quad \text{and} \quad |\det \nabla\mathcal{V}(X_0)| \geq \frac{C_1}{2}.$$

By the inverse function theorem there are positive constants, r_1 and r_2, independent of $\mathcal{V}$, such that $B_{r_1}(X_0) \subset \Omega$, $\mathcal{V}$ is a diffeomorphism from $B_{r_1}(X_0)$ onto its image, and $B_{r_2}(\mathcal{V}(X_0)) \subset \mathcal{V}(B_{r_1}(X_0))$. Setting $\varepsilon_2 = \min(\varepsilon_1, r_2/2)$, we have

$$0 = \mathcal{U}(X_0) \in B_{r_2}(\mathcal{V}(X_0)) \subset \mathcal{V}(B_{r_1}(X_0)) \quad \text{if } \mathcal{V} \in \mathcal{S}_{\varepsilon_2} \quad \text{and} \quad \|\mathcal{U} - \mathcal{V}\|_{C^2(\overline{\Omega})} < \varepsilon_2.$$

Next we note that there exists $\eta > 0$ such that

$$\inf_{X \in \Omega - B_{r_1}(X_0)} |\mathcal{U}(X)| \geq \eta \quad \text{for all } \mathcal{U} \text{ in } \mathcal{S}, \tag{3.10}$$

where $\{X_0\} = \Gamma(\mathcal{U})$. Indeed, if (3.10) fails then by compactness there is a member of $\mathcal{S}$ with more than one zero.

Now set $\varepsilon = \min(\varepsilon_2, \eta/2)$. If $\mathcal{V} \in \mathcal{S}_\varepsilon$, it follows that $\mathcal{V} \neq 0$ in $\Omega - B_{r_1}(X_0)$ whenever $\mathcal{U} \in \mathcal{S}$, $\|\mathcal{V} - \mathcal{U}\|_{C^2(\overline{\Omega})} < \varepsilon$ and $\{X_0\} = \Gamma(\mathcal{U})$. Moreover, $\mathcal{V}|_{B_{r_1}(X_0)}$ is a diffeomorphism and $0 \in \mathcal{V}(B_{r_1}(x_0))$. □

For the remainder of this section, we assume (in addition to our hypotheses on ψ_0) that $\mathcal{U}_0 \in \mathcal{M}$ and $\mathcal{U}(X,t)$ is a solution of (3.1).

Let $\mathcal{U}(t) = \mathcal{U}(\cdot, t)$. Our next result states that if $J(\mathcal{U}_0)$ is sufficiently close to the minimum energy level, then $\Gamma(\mathcal{U}(t))$ is a single point for all t sufficiently large.

THEOREM 3.3. *Let $m = \inf_{\psi \in \mathcal{M}} J(\psi)$. There exist positive constants δ and T (depending only on Ω, ψ_0, and $\mathcal{U}_0$) such that if $J(\mathcal{U}_0) \leq m + \delta$, then $\Gamma(\mathcal{U}(t)) = \{X(t)\}$ for all $t \geq T$, where $X(t)$ is a C^1 curve from (T, ∞) into Ω.*

Proof. Let

$$\omega = \omega(\mathcal{U}_0) \equiv \{\mathcal{V} : \mathcal{U}(t_n) \to \mathcal{V} \text{ in } C^2(\overline{\Omega}) \text{ for some sequence } t_n \to \infty\}.$$

By Theorem 3.1, $\{\mathcal{U}(t) : t \geq 1\}$ is precompact in $C^2(\overline{\Omega})$. Thus ω is nonempty and for each $\varepsilon > 0$, there exists $T(\varepsilon) > 0$ such that $\{\mathcal{U}(t) : t \geq T(\varepsilon)\} \subset \omega_\varepsilon$. Multiplying (3.1) by $\mathcal{U}_t(t)$ and integrating gives

$$\int_0^t \int_\Omega |\mathcal{U}_t|^2 dX\, dt + \frac{1}{2} J(\mathcal{U}(t)) = \frac{1}{2} J(\mathcal{U}_0). \tag{3.11}$$

From this and the argument of Theorem 1.1 in [L-P], it follows that $\omega(\mathcal{U}_0)$ is made up of solutions to (2.1).

Given $\delta > 0$ assume that $J(\mathcal{U}_0) \leq m + \delta$. From (3.11), $J(\mathcal{V}) \leq m + \delta$ for any $\mathcal{V}$ in $\omega(\mathcal{U}_0)$. By (2.2) and compactness, it follows that for each $\varepsilon > 0$ there exists $\delta_1 > 0$ such that if $\mathcal{V}$ solves (2.1) and $J(\mathcal{V}) \leq m+\delta_1$, then $\mathcal{V} \in \mathcal{S}_\varepsilon$. Consequently, if $J(\mathcal{U}_0) \leq m+\delta_1$, there exists $T = T(\Omega, \mathcal{U}_0, \varepsilon)$ such that if $t \geq T$, then $\mathcal{U}(t) \in \mathcal{S}_{2\varepsilon}$. By Lemma 3.2 if $t \geq T$, we have $\Gamma(\mathcal{U}(t)) = \{X(t)\}$ for some $X(t)$ in Ω and $\det \nabla\mathcal{U}(X(t),t) \neq 0$. Since $\mathcal{U}(X(t),t) = 0$, the implicit function theorem ensures that $X(t)$ is C^1 and

$$\frac{dX}{dt} = -[\nabla\mathcal{U}(X(t),t)]^{-1}\frac{\partial\mathcal{U}}{\partial t}(X(t),t). \qquad \square$$

A result of Simon [S] implies that if $G(\tau)$ is real analytic for $\tau \geq 0$ and if ψ_0 and Ω are sufficiently regular, then $\mathcal{U}(t)$ converges to an equilibrium solution, $\mathcal{U}^\infty$ as $t \to \infty$. We use this result to prove that $X(t) \to X^\infty$ as $t \to \infty$, where $\{X^\infty\} = \Gamma(\mathcal{U}^\infty)$. (See Theorem 3.6.)

Simon's work is done in the Sobolev space $W^{k,2}(\Omega)$, where k is large enough so that

$$\|\mathcal{V}\|_{C^2(\overline{\Omega})} \leq C \cdot \|\mathcal{V}\|_{W^{k-3,2}(\Omega)} \quad \text{for all } \mathcal{V} \text{ in } W^{k-3,2}(\Omega).$$

Since $\Omega \subset \mathbb{R}^2$, it suffices to take $k = 7$.

LEMMA 3.4. *Assume that $G(\tau)$ is real analytic for $\tau \geq 0$, Ω is of class $C^{7,\mu}$, and $\psi_0 \in C^{7,\mu}(\overline{\Omega};\mathbb{R}^2)$. There exists a solution, $\mathcal{U}^\infty$, of* (2.1) *such that $\mathcal{U}(t) \to \mathcal{U}^\infty$ in $C^2(\overline{\Omega})$ as $t \to \infty$.*

Proof. We apply Corollary 1 from [S]. This asserts that $\mathcal{U}(t) \to \mathcal{U}^\infty$ in $W^{7,2}(\Omega)$ as $t \to \infty$, provided it is known that for some sequence, $\{t_n\}$, with $t_n \to \infty$, we have $\mathcal{U}(t_n) \to \mathcal{U}^\infty$ in $W^{7,2}(\Omega)$ and $\mathcal{U}^\infty$ solves (2.1). To prove this condition, note that by Theorem 3.1, $\{\mathcal{U}(t) : t \geq 1\}$ is precompact in $C^7(\overline{\Omega})$; hence it is precompact in $W^{7,2}(\Omega)$. In addition, the proof of Lemma 3.3 ensures that any limit point of this set is a solution of (3.1). □

A further result of Simon is that the only equilibrium solution near a minimizer is another minimizer. From this we obtain the following.

LEMMA 3.5. *Under the hypotheses of Lemma* 3.4, *there exists $\delta > 0$ such that if $\mathcal{U}_0 \in \mathcal{M}$ and $J(\mathcal{U}_0) \leq m + \delta$, then $\mathcal{U}^\infty(X) \equiv \lim_{t\to\infty}\mathcal{U}(X,t)$ is a minimizer.*

Proof. Assume that no such δ exists. Then there are sequences, $\{\mathcal{U}_{n0}(X)\}$ and $\{\mathcal{U}_n(X,t)\}$, such that $\mathcal{U}_{n0} \in \mathcal{M}$, $\mathcal{U}_n(X,t)$ satisfies (3.1) with $\mathcal{U}_0(X) = \mathcal{U}_{n0}(X)$,

$$J(\mathcal{U}_{n0}) \leq m + \frac{1}{n} \quad \text{and} \quad m < J(\mathcal{U}_n^\infty) \quad \text{for } n = 1, 2, \ldots,$$

where $\mathcal{U}_n^\infty(X) \equiv \lim_{t\to\infty}\mathcal{U}_n(X,t)$. By (2.2) and (3.11) we may assume (by considering a subsequence) that $\mathcal{U}_n^\infty \to \mathcal{V}$ in $C^{2,\mu/2}(\overline{\Omega})$, where $\mathcal{V}$ is a minimizer. By Theorem 3 of [S], there exists $\sigma > 0$ such that if

$$\|\mathcal{V} - \mathcal{U}_n^\infty\|_{C^{2,\mu/2}(\overline{\Omega})} < \sigma,$$

then $J(\mathcal{U}_n^\infty) = m$, which is impossible. □

From Theorem 3.3 and the above lemmas, we obtain the following.

THEOREM 3.6. *Assume that Ω is of class $C^{7,\mu}$, G is real analytic in* $[0,\infty)$, *and $\psi_0 \in C^{7,\mu}(\overline{\Omega};\mathbb{R}^2)$. There exists $\delta > 0$ and a point X^∞ in Ω so that if $J(\mathcal{U}_0) \leq m + \delta$, then*

(i) *$\mathcal{U}(t) \to \mathcal{U}^\infty$ in $C^2(\overline{\Omega})$ as $t \to \infty$;*

(ii) *$\Gamma(\mathcal{U}(t)) = \{X(t)\} \subset \Omega$ for all t sufficiently large;*

(iii) $X^\infty = \lim_{t\to\infty} X(t)$ *and* $\Gamma(\mathcal{U}^\infty) = \{X^\infty\}$.

Proof. By Lemmas 3.4 and 3.5, (i) holds for a minimizer, $\mathcal{U}^\infty$. Condition (ii) follows from Theorem 3.3, and then condition (iii) holds by continuity and Theorem 2.3. □

REFERENCES

[Ab] A. A. Abrikosov, Soviet Phys., 5 (1957), p. 1174.

[Al] N. D. Alikakos, *Quantitative maximum principles and strongly coupled gradient-like reaction-diffusion systems*, Prov. Royal Soc. Edinburgh Sect. A, 94 (1983), pp. 265–286.

[G] M. Giaquinta, *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Systems*, Princeton University Press, Princeton, NJ, 1983.

[G-L] V. L. Ginzburg and L. D. Landau, JETP, 20 (1950), p. 1064.

[Gr] E. P. Gross, *Dynamics of interacting bosons*, in Physics of Many-Particle Systems: Methods and Problems, Vol. 1, E. Meeron, ed., Gordon and Breach, New York, 1966.

[H-W] P. Hartman and A. Wintner, *On the local behavior of solutions of non-parabolic partial differential equations* (I), Amer. J. Math., 75 (1953), pp. 449–476.

[H] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, New York, 1981.

[L-S-U] O. A. Ladyẑenskaya, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Translations of Mathematical Monographs, Vol. 23, American Mathematical Society, Providence, RI, 1968.

[L-P] M. Langlais and D. Phillips, *Stabilization of solutions of non-linear and degenerate evolution equations*, Nonlinear Anal. Theory, Meth. Appl., 9 (1985), pp. 321–333.

[N] J. C. Neu, *Vortices in complex scalar fields*, Phys. D, 43 (1990), pp. 385–406.

[P] L. E. Payne, *On two conjectures in the fixed membrane eigenvalue problem*, J. Appl. Math. Phys., 24 (1973), pp. 721–729.

[S] L. Simon, *Asymptotics for a class of nonlinear evolution equations, with applications to geometric problems*, Ann. Math., 118 (1983), pp. 525–571.

SIAM J. MATH. ANAL.
Vol. 24, No. 5, pp. 1294–1311, September 1993

010

SUBHARMONIC SOLUTIONS FOR SOME SECOND-ORDER DIFFERENTIAL EQUATIONS WITH SINGULARITIES*

ALESSANDRO FONDA[†], RAÚL MANÁSEVICH[‡], AND FABIO ZANOLIN[§]

Abstract. The existence of infinitely many subharmonic solutions is proved for the periodically forced nonlinear scalar equation $u'' + g(u) = e(t)$, where g is a continuous function that is defined on a open proper interval $(A, B) \subset \mathbb{R}$. The nonlinear restoring field g is supposed to have some singular behaviour at the boundary of its domain. The following two main possibilities are analyzed:

(a) The domain is unbounded and g is sublinear at infinity. In this case, via critical point theory, it is possible to prove the existence of a sequence of subharmonics whose amplitudes and minimal periods tend to infinity.

(b) The domain is bounded and the periodic forcing term $e(t)$ has minimal period $T > 0$. In this case, using the generalized Poincaré–Birkhoff fixed point theorem, it is possible to show that for any $m \in \mathbb{N}$, there are infinitely many periodic solutions having mT as minimal period.

Applications are given to the dynamics of a charged particle moving on a line over which one has placed some electric charges of the same sign.

Key words. periodic solutions, subharmonics, repulsive singularities, saddle point theorem, critical levels, twist maps, generalized Poincaré–Birkhoff theorem

AMS subject classifications. 34C15, 34C25, 58E05, 70D05

1. Introduction.

A scalar equation of the form

$$u'' + g(u) = e(t) \tag{1.1}$$

can be viewed as a model for a system with one degree of freedom subject to an internal force given by the nonlinear restoring field $g(u)$ and an external time-dependent perturbation represented by $e(t)$.

In this paper we are interested in situations where $g(u)$ is a field having one or more singularities, all of which are of repelling type, and $e(t)$ will be supposed to be a periodic forcing, with period $T > 0$. We prove the existence of subharmonic solutions of (1.1), i.e., periodic solutions whose periods are integer multiples of T.

A simple physical model for this type of equation can be given by the dynamics of a charged particle moving on a line, over which one has placed some electric charges of the same sign. Since we consider only trajectories which do not collide with the singularity points, we can reduce our study to two different cases: the case of one singularity, with the particle moving on one side, and the case of two singularities, with the particle in between.

In §2 we deal with the one-singularity case. This case has been already considered by Lazer and Solimini in [17] (see also [14], [18]). They proved the existence of at least

*Received by the editors February 19, 1992; accepted for publication (in revised form) November 30, 1992.

[†] Dipartimento di Scienze Matematiche, Università di Trieste, P. le Europa 1, 34127, Trieste, Italy. The research of this author was sponsored by MURST 60%.

[‡] Departamento de Ingenieria Matemáticas, F.C.F.M. Universidad de Chile, Casilla 170, Correo 3, Santiago, Chile. The research of this author was sponsored by FONDECYT (1209-91) and by the European Economic Community.

[§] Dipartimento di Matematica e Informatica, Università di Udine, Via Zanon 6 - 33100, Udine, Italy. The research of this author was sponsored by the European Economic Community.

one T-periodic solution of (1.1) for the model described above. Under some general assumptions on g, we will show that, besides the T-periodic solutions, there is a whole sequence of subharmonic solutions whose minimal period is an arbitrarily large integer multiple of T. The proofs of the results of this section will use variational arguments providing critical points of saddle type for the action functional.

The forced two-singularities case for (1.1), as far as we know, has not been considered explicitly in the literature. Some work has been done for systems with a potential well (see [2], [1], [4]), but without a forcing term.

In §3, we will consider this case. Using a generalized version of the Poincaré–Birkhoff fixed point theorem we will prove that, for any fixed period which is an integer multiple of T, there are infinitely many periodic solutions with such a minimal period.

To be more specific, we will consider $g(u)$ to be defined on an open interval (A, B), which may be bounded or unbounded. This will permit us to deal simultaneously with various different qualitative situations. Intuitively, we may think of the extreme points A and B of the domain of g as "singularities" for the field g. With this in mind, it is reasonable to look for conditions on g such that g grows faster than linear at the singularities. Such requirement is satisfied when A (respectively, B) is finite, and $\lim_{x\to A^+} g(x) = -\infty$ (respectively, $\lim_{x\to B^-} g(x) = +\infty$) while, for $A = -\infty$ (respectively, $B = +\infty$), we will assume that $g(x)/x \to +\infty$ as $x \to A^+$ (respectively, $x \to B^-$).

In this setting, the search of T-periodic solutions and subharmonic solutions in the case $A = -\infty$ and $B = +\infty$ has already been considered in several papers starting with Morris [19], [20] who proved in [20] the existence of infinitely many subharmonics of any order for e smooth and $g(x) = 2x^3$. Extensions of Morris's result were obtained in [9], [7] for any g continuous and such that $g(x)/x \to +\infty$ for $x \to \pm\infty$, using a generalization of the Poincaré–Birkhoff fixed point theorem due to W. Ding [9]. Namely, the existence of fixed points for the iterates of the Poincaré map associated to (1.1) is obtained in [8], by showing that there are circular annuli in the plane (u, u') where the twist condition at the boundaries (which are circumferences) is satisfied.

In [5] Del Pino and Manásevich considered the case $A \in \mathbb{R}$ and $B = +\infty$ for a variant of (1.1) motivated by a problem in nonlinear elasticity. They proved the existence of infinitely many T-periodic solutions using the more refined version of W. Ding's theorem in [10], where fixed points of an area-preserving homeomorphism twisting the boundaries of an annulus are obtained for annuli with star-shaped boundaries. Note that in this case, the singularity in A modifies the geometry of the planar flow and now the twist property has to be checked on the boundary of some annular regions which are "deformations" of circular annuli through a non-Euclidean metric. For another recent application of the Poincaré–Birkhoff theorem to (1.1), see also [6].

In §4, we apply our results to the dynamics of an electric charge moving in a Coulombian field with one or two singularities.

2. Sublinear case and one singularity.

We consider the equation

$$u'' + g(t, u) = e(t), \tag{2.1}$$

where $g : \mathbb{R} \times (0, +\infty) \to \mathbb{R}$ is a Caratheodory function, T-periodic in its first variable, such that for every positive constants $r < R$ there is a $\nu = \nu_{r,R} \in L^1(0, T)$ with $|g(t, x)| \le \nu(t)$ for almost every $t \in [0, T]$ and all $x \in [r, R]$. Moreover, $e : \mathbb{R} \to \mathbb{R}$ is locally integrable and T-periodic $(T > 0)$. We denote by $\bar{e}$ the mean value of $e(t)$, i.e., $\bar{e} = (1/T) \int_0^T e(t)\, dt$.

THEOREM 2.1. *Let $F : (0,+\infty) \to \mathbb{R}$ be a continuously differentiable function satisfying the following two properties:*

$$\lim_{x\to+\infty} \frac{F(x)}{x^2} = 0, \tag{k$_1$}$$

$$\lim_{x\to 0^+} F(x) = +\infty. \tag{k$_2$}$$

Assume that

$$g(t,x) \le F'(x), \tag{k$_3$}$$

and

$$g(t,x)\operatorname{sgn}(x-1) \ge -h(t), \tag{k$_4$}$$

for all $x > 0$ and almost every $t \in [0,T]$, where $h \in L^1([0,T],\mathbb{R}^+)$. If, moreover,

$$\frac{1}{T}\int_0^T \limsup_{x\to 0^+} g(t,x)\, dt \ < \bar{e} < \frac{1}{T}\int_0^T \liminf_{x\to+\infty} g(t,x)\, dt, \tag{k$_5$}$$

then (2.1) has a sequence $(x_k)_{k\ge 1}$ of positive kT-periodic solutions whose minimal periods tend to infinity.

We first define a truncation function. Thus, for $r > 0$, let us define $g_r : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ as follows:

$$g_r(t,x) = \begin{cases} g(t,x) & \text{if } x \ge r; \\ g(t,r) & \text{if } x < r. \end{cases}$$

PROPOSITION 2.1. *For every $k \in \mathbb{N}$ there exist positive r_k, R_k such that for any $s \in (0, r_k]$ and any kT-periodic solution u of*

$$u'' + g_s(t,u) = e(t), \tag{2.2}$$

we have that $r_k \le u(t) \le R_k$ for all $t \in \mathbb{R}$. In particular, any kT-periodic solution of (2.2) with $s = r_k$ is a solution of (2.1).

Proof. Without loss of generality, we can assume that $\bar{e} = 0$ (just subtract $\bar{e}$ to both sides of (2.2)).

We argue by contradiction. Fix $k \in \mathbb{N}$ and assume that for every $n \in \mathbb{N}$, there are $s_n \in (0, 1/n)$ and a kT-periodic function u_n that satisfy

$$u_n'' + g_{s_n}(t,u_n) = e(t), \tag{2.3}$$

and such that $\{u_n(t) \mid t \in \mathbb{R}\} \not\subset [1/n, n]$.

In the following, we denote by $||\cdot||_q$ the usual L^q-norm on $(0,kT)$.

We claim there exists a $d \ge 1$ such that for every n there is $t_n^{(1)} \in [0,kT]$ with $u_n(t_n^{(1)}) \in [1/d, d]$.

Indeed, suppose for instance that, for a subsequence, $\max u_n \to c \in [-\infty, 0]$. Since

$$0 = \int_0^{kT} g_{s_n}(t, u_n(t))\, dt, \tag{2.4}$$

by Fatou's lemma we have

$$0 \leq \int_0^{kT} \limsup_{n\to\infty}[g(t,u_n(t))\chi_{[u_n>s_n]} + g(t,s_n)\chi_{[u_n\leq s_n]}]\ dt \ \leq \int_0^{kT} \limsup_{x\to 0^+} g(t,x)\ dt$$

contradicting the Landesman-Lazer like condition (k_5). A similar contradiction is obtained if we let $\min u_n \to +\infty$, and the claim follows.

From now on we fix the constant $d \geq 1$ such that (according to (k_3) and (k_5)), $F'(x) > 0$ for each $x \in [d, +\infty)$.

Next, let us prove that there exists a positive constant R such that $\max u_n \leq R$ for every n. We will use some ideas from [22]. By contradiction, assume there exists a subsequence, still denoted by (u_n) for which $\max u_n \to +\infty$. Then we can find an interval $[\alpha_n, \beta_n]$, containing a point $t_n^{(2)}$ with $u_n(t_n^{(2)}) = \max u_n$, such that $(\beta_n - \alpha_n) \leq kT$ and

$$u_n(\alpha_n) = d = u_n(\beta_n);$$
$$d \leq u_n(t) \leq u_n(t_n^{(2)}) \quad \text{for all } t \in [\alpha_n, \beta_n].$$

For $t \in [\alpha_n, \beta_n]$, we have that (2.3) can be written as

$$\begin{aligned} u_n' &= v_n + \int_{\alpha_n}^t e(s)\ ds \\ v_n' &= -g(t, u_n). \end{aligned} \tag{2.5}$$

Since $v_n(t) - \int_{\alpha_n}^t h(s)\ ds$ is decreasing in this interval, using (2.5) we obtain

$$\max u_n - d \leq kT(v_n(\alpha_n) + ||e||_1 + ||h||_1), \tag{2.6}$$

so that, for n large enough, $v_n(\alpha_n) > ||e||_1$. On the other hand, again from (2.5) we find that $v_n(t_n^{(2)}) \leq ||e||_1$. Thus there exists a $t_n^{(3)} \in (\alpha_n, t_n^{(2)}]$ such that $v_n(t_n^{(3)}) = ||e||_1$. For t in the interval $[\alpha_n, t_n^{(3)}]$ we have

$$\begin{aligned} &\frac{d}{dt}\left[F(u_n(t)) + \frac{1}{2}(v_n(t) - ||e||_1)^2\right] \\ &\quad = F'(u_n(t))\left[v_n(t) + \int_{\alpha_n}^t e(s)\ ds\right] + (v_n(t) - ||e||_1)(-g(t,u_n(t))) \\ &\quad \geq F'(u_n(t))(v_n(t) - ||e||_1) + (v_n(t) - ||e||_1)(-g(t,u_n(t))) \\ &\quad \geq (F'(u_n(t)) - g(t,u_n(t)))[v_n(t) - ||e||_1] \geq 0. \end{aligned}$$

Thus $F(u_n(\cdot)) + \frac{1}{2}(v_n(\cdot) - ||e||_1)^2$ is increasing in this interval and hence

$$F(d) + \tfrac{1}{2}(v_n(\alpha_n) - ||e||_1)^2 \leq F(u_n(t_n^{(3)})).$$

From assumption (k_1) we find that for any $\varepsilon > 0$ there is $C'_\varepsilon > 0$ such that

$$F(u) \leq \varepsilon u^2 + C'_\varepsilon \quad \text{for every} \quad u \geq d.$$

Hence

$$\begin{aligned} F(d) + \tfrac{1}{2}(v_n(\alpha_n) - ||e||_1)^2 &\leq \varepsilon(u_n(t_n^{(3)}))^2 + C'_\varepsilon \\ &\leq \varepsilon(\max u_n)^2 + C'_\varepsilon. \end{aligned}$$

Now choosing ε small enough and calling on (2.6) we obtain a contradiction when $n \to \infty$. Thus, we have proved that there exists $R > 0$ such that $\max u_n \leq R$ for every n.

Next, from (2.4) and $s_n \leq 1/d$ we have

$$\begin{aligned}\int_{[u_n<1/d]} |g_{s_n}(t,u_n(t))|\,dt &\leq \int_{[u_n<1/d]} (-g_{s_n}(t,u_n(t))+h(t))\,dt \;+||h||_1\\ &\leq \int_{[1/d\leq u_n\leq R]} |g(t,u_n(t))|\,dt \;+2||h||_1 \;\leq\; C,\end{aligned}$$

and we obtain that $||g_{s_n}(\cdot,u_n(\cdot))||_1 \leq 2C$; hence

$$||u_n'||_\infty \leq C_1 := 2C + ||e||_1.$$

Now define $\tilde{g}_{s_n}(t,x) := g_{s_n}(t,x) - h(t)$. Then (2.3) can be written as

$$u_n'' + \tilde{g}_{s_n}(t,u_n) = e(t) - h(t). \tag{2.7}$$

Set

$$f_{s_n}(x) = \begin{cases} F'(x) & \text{if } x \geq s_n \\ F'(s_n) & \text{if } x < s_n, \end{cases}$$

and

$$\eta_{s_n}(x) = \min\,\{0, f_{s_n}(x)\}.$$

Since $h(t) \geq 0$, we have that, for $x \leq 1$, $\tilde{g}_{s_n}(t,x) \leq \eta_{s_n}(x)$. Assume next there is a $t_n^{(4)} > t_n^{(1)}$ such that $u(t_n^{(4)}) < 1/n$. Then there are $t_n^{(5)} < t_n^{(6)}$ such that $[t_n^{(5)}, t_n^{(6)}]$ is contained in $[t_n^{(1)}, t_n^{(4)}]$, and such that $u_n(t_n^{(5)}) = 1/d$, $u_n(t_n^{(6)}) = 1/n$ and $1/n \leq u_n(t) \leq 1/d$ for all $t \in [t_n^{(5)}, t_n^{(6)}]$. Note that $t_n^{(6)} - t_n^{(5)} \leq kT$.

Then, multiplying (2.7) by $(u_n' - C_1)$ and integrating over $[t_n^{(5)}, t_n^{(6)}]$, we get

$$\begin{aligned}\frac{1}{2}(u_n'(t_n^{(6)}) - C_1)^2 - \frac{1}{2}(u_n'(t_n^{(5)}) - C_1)^2 + \int_{t_n^{(5)}}^{t_n^{(6)}} \tilde{g}_{s_n}(t,u_n(t))(u_n'(t) - C_1)\,dt&\\ \leq\; 2C_1(||e||_1 + ||h||_1).&\end{aligned}$$

Thus, since $(u_n' - C_1) \leq 0$ and $\eta_{s_n}(u_n(t)) \leq 0$, we obtain

$$\int_{t_n^{(5)}}^{t_n^{(6)}} \eta_{s_n}(u_n(t))u_n'(t)\,dt \leq \int_{t_n^{(5)}}^{t_n^{(6)}} \eta_{s_n}(u_n(t))(u_n'(t) - C_1)\,dt \leq \tilde{C},$$

where $\tilde{C} := 2C_1(C_1 + ||e||_1 + ||h||_1)$. Setting $H_{s_n}(x) = \int_{1/d}^{x} \eta_{s_n}(\xi)\,d\xi$ (a primitive of η_{s_n}), it follows that

$$H_{s_n}(1/n) = H_{s_n}(u_n(t_n^{(6)})) - H_{s_n}(u_n(t_n^{(5)})) \leq \tilde{C}.$$

But

$$H_{s_n}(1/n) \geq \int_{1/d}^{1/n} F'(x)\,dx \geq F(1/n) - F(1/d) \to +\infty \quad \text{as} \quad n \to \infty,$$

and hence we have a contradiction. Thus the proposition is proved. □

Proof of Theorem 2.1. We consider (2.2) with $s = r_k$, i.e.,

$$u'' + g_{r_k}(t,u) = e(t). \tag{2.8}$$

At this point we are in the same situation as in the proof of Theorem 2.5 in [12]. This is why we prefer to give only the main lines of the proof, the details being available from [12]. We take r_k sufficiently small, so that $(1/T)\int_0^T g(t,r_k)\,dt < \bar{e}$, and in such a way that $r_k \to 0$ as $k \to \infty$. Denoting by G_{r_k} a primitive of g_{r_k} with respect to its second variable, we have that for every k we are in the situation of [12, Thm. 2.1]. Thus we can apply the Saddle Point theorem to the functional $\phi_k : H^1_{kT} \to \mathbb{R}$ defined by

$$\phi_k(u) = \int_0^{kT} \left(\frac{1}{2}(u')^2 - G_{r_k}(t,u) + eu\right)\,dt,$$

whose critical points correspond to the kT-periodic solutions of (2.8). We find that for each k there is a $\rho_k > 0$ sufficiently large and a critical point u_k of ϕ_k such that

$$\phi_k(u_k) = \inf_{\gamma\in\Gamma_k}\max_{\xi\in[-\rho_k,\rho_k]} \phi_k(\gamma(\xi)),$$

where $\Gamma_k = \{\gamma \in C([-\rho_k,\rho_k],H^1_{kT}) \mid \gamma(\pm\rho_k) = \pm\rho_k\}$. By Fatou's lemma, we have

$$\liminf_{|x|\to\infty,k\to\infty} \operatorname{sgn}(x)\int_0^T\int_0^1 g_{r_k}(t,xs)\,ds\,dt > \bar{e}\,T,$$

and since $\int_0^T G_{r_k}(t,x)\,dt = x\int_0^T\int_0^1 g_{r_k}(t,xs)\,ds\,dt$,

$$\liminf_{|x|\to\infty,k\to\infty} \int_0^T G_{r_k}(t,x)\,dt \; - \bar{e}xT = +\infty.$$

Reasoning next as in the proof of [12, Thms. 2.1 and 2.5] (see also [11]), we can show that

$$\lim_{k\to\infty} \frac{1}{k}\phi_k(u_k) = -\infty. \tag{2.9}$$

Now we can prove that the minimal periods of the kT-periodic solutions u_k tend to infinity as $k \to \infty$. If not, for a subsequence there would be a subsequence with a common period, say $\bar{k}T$. Noting that from Proposition 2.1 the set of $\bar{k}T$-periodic solutions of (2.8) is bounded in H^1_{kT}, independently of $k \geq \bar{k}$, we get a contradiction with (2.9). □

As a consequence of Theorem 2.1 we have the following (cf. [12, Cor. 2.6]).

COROLLARY 2.1. *Suppose* $g : \mathbb{R} \times (0,+\infty) \to \mathbb{R}$ *to be continuous, and that*

$$\lim_{x\to+\infty} \frac{g(t,x)}{x} = 0, \tag{m$_1$}$$

$$\limsup_{x\to0^+} xg(t,x) \leq c < 0, \tag{m$_2$}$$

uniformly with respect to t. *If moreover*

$$\bar{e} < \frac{1}{T}\int_0^T \liminf_{x\to+\infty} g(t,x)\,dt, \tag{m$_3$}$$

the conclusion of Theorem 2.1 *holds.*

If we substitute (k_4)–(k_5) with the more restrictive sign condition

$$\limsup_{x\to 0^+} g(t,x) \le c_1 < \bar{e} < c_2 \le \liminf_{x\to+\infty} g(t,x), \tag{k$'_5$}$$

uniformly with respect to t, it is possible to get more precise information about the subharmonic solutions of (2.1), as follows.

THEOREM 2.2. *Assume* (k_1), (k_2), (k_3), *and* (k'_5). *Then there exists an integer* $m^* \in \mathbb{N}$ *such that for every* $m \ge m^*$, *equation* (2.10) *has at least one positive periodic solution having minimal period* mT.

The proof of this statement is omitted since it can be achieved via the generalized Poincaré–Birkhoff fixed point theorem arguing as in [8] (proof of Theorem 1.1). On the other hand, we prefer to present a different application of this theorem to the case of two singularities in the next section.

Remark 2.1. It is possible to see that, in Theorem 2.1, we can replace conditions (k_4) and (k_5) by

(k''_4) there is a $d \ge 1$ such that $g(t,x)\mathrm{sgn}(x-1) > \bar{e}$ for all $x \in (0,1/d) \cup (d,+\infty)$;

$$\lim_{x\to+\infty}\int_0^T (G(t,x) - \bar{e}x)\,dt = +\infty; \tag{k$''_5$}$$

where G is a primitive of g with respect to the x variable. We are then led to the following.

COROLLARY 2.2. *Assume that* $g(t,x) = g(x)$ *and that the following conditions hold:*

$$\lim_{x\to+\infty}\frac{G(x)}{x^2} = 0; \tag{j$_1$}$$

(j_2) *There is a* $d \ge 1$ *such that* $g(x)\mathrm{sgn}(x-1) > \bar{e}$ *for all* $x \in (0,1/d) \cup (d,+\infty)$;

$$\lim_{x\to 0^+} G(x) = \lim_{x\to+\infty}(G(x) - \bar{e}x) = +\infty. \tag{j$_3$}$$

Then the same conclusion of Theorem 2.1 *holds.*

Remark 2.2. By a suitable change of variables, we can easily restate the analogous version of the results of this section in the case where $g(t,x)$ is defined on $\mathbb{R}\times(A,+\infty)$ or on $\mathbb{R}\times(-\infty,B)$, the singularity being at $A \in \mathbb{R}$ or at $B \in \mathbb{R}$, respectively.

3. Superlinear case and two singularities. Consider again equation

$$u'' + g(u) = e(t), \tag{3.1}$$

where $g : (A,B) \to \mathbb{R}$ is continuous and $e : \mathbb{R} \to \mathbb{R}$ is T-periodic ($T > 0$), with $e \in L^1_{\mathrm{loc}}$.

Here we suppose that

$$-\infty \le A < B \le +\infty$$

and fix any $c \in (A,B)$.

Our goal is to prove the existence of infinitely many subharmonics of any order for (3.1) with $(u(t), u'(t))$ lying in the open strip

$$\mathcal{S} := (A, B) \times \mathbb{R}$$

and giving a precise statement about the nodal properties of $u(t) - c$.

To this end we consider also the equivalent system

$$u' = v + E(t), \qquad v' = -g(u) + \bar{e}, \tag{3.2}$$

where $E(t) := \int_0^t \left(e(\xi) - \bar{e}\right) d\xi$ and $\bar{e} := (1/T) \int_0^T e(t)\, dt$. Note that $E : \mathbb{R} \to \mathbb{R}$ is continuous and T-periodic with mean value zero. Let G be a primitive of g; e.g., the one defined by

$$G(x) := \int_c^x g(s)\, ds.$$

To describe our result, we assume further the next conditions.

$$\lim_{x \to A^+} G(x) = \lim_{x \to B^-} G(x) = +\infty. \tag{i$_1$}$$

$$\lim_{x \to A^+} \frac{g(x)}{(x - c)} = \lim_{x \to B^-} \frac{g(x)}{(x - c)} = +\infty. \tag{i$_2$}$$

Remark 3.1. Note that condition (i$_2$) reads as follows:

If $A = -\infty$, then $\lim_{x \to -\infty} g(x)/x = +\infty$, while if $A \in \mathbb{R}$, then $\lim_{x \to A^+} g(x) = -\infty$.

If $B = +\infty$, then $\lim_{x \to +\infty} g(x)/x = +\infty$, while if $B \in \mathbb{R}$, then $\lim_{x \to B^-} g(x) = +\infty$.

We also observe that (i$_1$) is always satisfied at $A = -\infty$ or at $B = +\infty$ when (i$_2$) is assumed.

Finally, we remark that (i$_2$) is independent upon the choice of the point $c \in (A, B)$.

We further introduce the following terminology (see, e.g., [15, p.17]).

For $f : \mathcal{O} \to \mathcal{S}$, $\alpha \mapsto f(\alpha)$, with $\alpha \in \mathcal{O}$, where $(\mathcal{O}, \prec)$ is a directed set, we write

$$f(\alpha) \to \partial\mathcal{S}$$

if, for every compact set $\mathcal{K} \subset \mathcal{S}$, there is $\alpha_{\mathcal{K}} \in \mathcal{O}$ such that $f(\alpha) \notin \mathcal{K}$, for all $\alpha \in \mathcal{O}$ with $\alpha \succ \alpha_{\mathcal{K}}$.

The case in which $e(t)$ is a constant, and hence $E(t) \equiv 0$, can be completely analyzed in terms of energy levels arguments. Indeed, we can prove that if $\Gamma(\mathrm{z}_0)$ is the orbit of

$$u' = v, \qquad v' = -g(u) + \bar{e},$$

with $(u(0), v(0)) = \mathrm{z}_0$, then, for $\mathrm{z}_0 \to \partial\mathcal{S}$, $\Gamma(\mathrm{z}_0)$ is a periodic orbit with minimal period tending to zero. Therefore we assume henceforth that $e(t) \not\equiv \bar{e}$ on a set of positive measure, so that $E(t)$ is nonconstant and it has a positive minimal period. Without loss of generality, we can suppose that T is the minimal period of $E(t)$.

Now we are in position to state our main result.

THEOREM 3.1. *Assume* (i$_1$) *and* (i$_2$) *and let* $m \geq 1$ *be any fixed integer. Then* (3.1) *has infinitely many periodic solutions with minimal period* mT. *More precisely, for each* $m \geq 1$, *there is an integer* $\nu_m^* \geq 0$ *such that for every* $p \in \mathbb{N}$ *with* p *prime with* m *and* $p > \nu_m^*$, *equation* (3.1) *has at least one periodic solution* $u = u_{m,p}(\cdot)$, *with minimal period* mT *and such that* $u(t) - c$ *has exactly* $2p$ *simple zeros in the interval* $[0, mT)$. *Moreover,* $(u_{m,p}(t), u'_{m,p}(t)) \to \partial S$, *as* $p \to +\infty$, *uniformly with respect to* $t \in [0, mT]$.

Remark 3.2. Note that if we are interested only in the existence of T-periodic solutions, then we can apply the above theorem taking $m = 1$, and we find ν_1^* such that for every $p \in \mathbb{N}$ with $p > \nu_1^*$ there is at least one T-periodic solution $u_p(\cdot)$ with $u_p(t) - c$ having exactly $2p$ simple zeros in the interval $[0, T)$.

This remark is true even in the case when T is not the minimal period of $E(t)$; however, in such a situation, we cannot guarantee that the periodic solutions we find have T as minimal period.

The proof of Theorem 3.1 is based on W. Ding's generalized Poincaré–Birkhoff fixed point theorem [10] which provides fixed points for the Poincaré map (and its mth iterates) associated to system (3.2). To do this, we need to have such an operator well defined. Hence, a first requirement is to have the uniqueness of the solutions for the Cauchy problems associated to (3.2). This difficulty can be overcome by a standard smoothing of the field g as briefly described in [23]. Of course, then it will be necessary to prove that the fixed points related to the approximating equations are all contained in the same annulus in order to pick up a sequence of these fixed points converging to a fixed point representing the initial value of an mT-periodic solution to (3.2). Here we do not follow such a program which has been already accomplished with all the details in various preceding papers. Accordingly, from now on, *we assume the uniqueness of the solutions for the Cauchy problems associated to system* (3.2) leaving the interested reader to complete the missing details following, e.g., [7].

We also remark that if we assume condition (i$_1$)–(i$_2$) then we have that the same is satisfied for the function $g(x) - \bar{e}$. Hence, calling $g(x)$ what was written before as $g(x) - \bar{e}$ and $e(t)$ instead of $e(t) - \bar{e}$, from now on we can assume, without loss of generality, that

$$\text{(i}_0\text{)} \qquad \frac{1}{T}\int_0^T e(t)dt = 0$$

holds. Notice that system (3.2) takes now the form

$$\text{(3.3)} \qquad u' = v + E(t), \qquad v' = -g(u).$$

Our first step is to prove the global existence in the past and in the future of the solutions to (3.3). In this direction we have the following result which is proved under some more general conditions than in Theorem 3.1.

PROPOSITION 3.1. *Assume* (i$_0$), (i$_1$) *and suppose that there are constants* a, b *with* $A < a < b < B$ *such that*

$$\text{(3.4)} \qquad g(x) < 0 \quad \textit{for } A < x \leq a, \qquad g(x) > 0 \quad \textit{for } b \leq x < B.$$

Then any noncontinuable solution $z = (u, v)$ *of* (3.3) *is defined in* $(-\infty, +\infty)$.

Proof. From (i$_1$) it is clear that $G(x) \geq G_{\min} > -\infty$, for all $x \in (A, B)$.

Let $\eta : (A, B) \to \mathbb{R}$, be a C^1 function defined as follows (cf. [16, p.120]): $\eta(x) = -||E||_\infty$ for $x \in (A, a)$, $\eta(x) = ||E||_\infty$ for $x \in (b, B)$, and η increasing in $[a, b]$. We thus have that $0 \leq \eta'(x) \leq L_\eta$ for all x in (A, B), where L_η is a suitable constant depending on the function η.

Now, define $V : (A, B) \times \mathbb{R} \to \mathbb{R}$ by

$$V(x, y) := \tfrac{1}{2}(y + \eta(x))^2 + G(x) - G_{\min}.$$

It is clear that for any $D > 0$ there is a compact set $\mathcal{B}_D$ contained in $\mathcal{S}$ such that

$$(x, y) \in \mathcal{S} \setminus \mathcal{B}_D \quad \text{implies that} \quad V(x, y) > D.$$

Next, let $z(t) = (u(t), v(t))$ be a solution of (3.3) defined in a maximal interval (α, β), with $\alpha < t_0 < \beta$ and $z(t_0) \in \mathcal{S}$. For $t \in [t_0, \beta)$, we have

$$\begin{aligned}\frac{d}{dt}V(z(t)) =&[\eta'(u)(v + \eta(u)) + g(u)]u' + [v + \eta(u)]v' \\ =&\eta'(u)[v + \eta(u)]^2 + \eta'(u)[v + \eta(u)][-\eta(u) + E(t)] - g(u)[\eta(u) - E(t)].\end{aligned}$$

Noting that $-g(x)[\eta(x) - E(t)] \leq 0$ for all $x \in (A, a) \cup (b, B)$, we find that there is a constant $R_0 > 0$ such that $-g(x)[\eta(x) - E(t)] \leq R_0$ for all $x \in (A, B)$ and $t \in \mathbb{R}$.

Thus

$$\begin{aligned}\frac{d}{dt}V(z(t)) &\leq L_\eta[v + \eta(u)]^2 + 2||E||_\infty L_\eta|v + \eta(u)| + R_0 \\ &\leq \frac{3}{2}L_\eta[v + \eta(u)]^2 + R_1,\end{aligned}$$

where R_1 is a constant depending on $||E||_\infty$ and R_0. We obtain

$$\frac{d}{dt}V(z(t)) \leq 3L_\eta V(z(t)) + R_1.$$

We claim that β must be $+\infty$. Otherwise from the last inequality and Gronwall's lemma,

$$V(z(t)) \leq \left[\frac{R_1}{3L_\eta} + V(z(t_0))\right] \exp(3L_\eta(\beta - t_0)) := \text{constant} := R_2.$$

Then

$$z(t) \in \mathcal{B}_{R_2} \quad \text{for all} \quad t \in [t_0, \beta).$$

But, this contradicts the global continuation theorem and the claim is proved.

Global continuability to the left follows from the above argument and by changing $t - t_0$ for $t_0 - t$ in (3.3). □

According to Proposition 3.1 and the uniqueness assumption for the Cauchy problems, we have that any solution to (3.3) is uniquely defined on $(-\infty, +\infty)$ by its initial conditions.

Remark 3.3. It is obvious that hypothesis (i_2) implies the existence of suitable constants a and b with

$$A < a < c < b < B \tag{3.5}$$

such that (3.4) holds. *Henceforth* (3.4) *and* (3.5) (*as well as* (i_0)) *will be constantly assumed in connection with* (i_2).

We define now the compact set

$$\mathcal{M} := [a,b] \times [-||E||_\infty, ||E||_\infty]$$

and observe that $(c,0) \in \mathcal{M}$. A corollary of Proposition 3.1 is the following.

PROPOSITION 3.2. *Assume* (i_1) *and* (i_2). *Then for every compact set* $\mathcal{K} \subset \mathcal{S}$ *and each* $m \in \mathbb{N}$, *there is a compact set* $\mathcal{B} = \mathcal{B}(\mathcal{K}, m) \subset \mathcal{S}$ *with* $\mathcal{K} \subset \mathcal{B}$ *such that for each solution* $z = (u,v)$ *of system* (3.3), *the following inference holds:*

$$z(0) \notin \mathcal{B} \Rightarrow z(t) \notin \mathcal{K} \quad \forall t \in [-mT, mT].$$

In particular, for each $m \in \mathbb{N}$, *there is a compact set* $\mathcal{R}_m$ *with* $\mathcal{M} \subset \mathcal{R}_m \subset \mathcal{S}$ *such that*

$$z(0) \notin \mathcal{R}_m \Rightarrow z(t) \notin \mathcal{M} \quad \forall t \in [-mT, mT] \tag{3.6}$$

holds for any solution z *to* (3.3).

By the second part of Proposition 3.2 and (3.6) we have that given any solution $z(\cdot) = (u(\cdot), v(\cdot))$ of equation (3.3) with $z(0) \notin \mathcal{R}_m$, it follows that if $t \in [0, mT]$ is such that $u(t) - c = 0$, or, respectively, $v(t) = 0$, then $u'(t) \neq 0$, respectively, $v'(t) \neq 0$. Then, according to [5], we can define the *rotation number* $\psi_m(u)$ as

$$\psi_m(u) = k\pi + \lim_{t \to 0^+} \tan^{-1}\left(\frac{v(t)}{u(t) - c}\right) - \lim_{t \to mT^-} \tan^{-1}\left(\frac{v(t)}{u(t) - c}\right), \tag{3.7}$$

where k is the number of zeros of $u(t) - c$ in $(0, mT)$. Geometrically, $\psi_m(u)$ represents the total angle the vector from the origin to the point $(u(t) - c, v(t))$ describes as t goes from zero to mT, positive angles measured clockwise.

On the other hand, if $z(0) \notin \mathcal{R}_m$, then $z(t) \neq (c,0)$, for all $t \in [0, mT]$ and therefore we can use polar coordinates with center in $(c,0)$ to express $z(t)$ via Prüfer transformation as

$$u(t) = c + \rho(t)\cos\theta(t), \qquad v(t) = \rho(t)\sin\theta(t).$$

By standard facts,

$$-\theta'(t) = \frac{g(u(t))(u(t) - c) + v(t)^2 + E(t)v(t)}{(u(t) - c)^2 + v(t)^2} \tag{3.8}$$

and by the definition of the rotation number, we have

$$\theta(0) - \theta(mT) = \psi_m(u). \tag{3.9}$$

We recall that, from Proposition 3.1 and the uniqueness of the solutions to the Cauchy problems associated to system (3.3), we have that for every $\mathrm{z}_0 = (\mathrm{x}_0, \mathrm{y}_0) \in \mathcal{S}$ there is a unique solution $z(t) = z(t; \mathrm{z}_0) = \big(u(t; \mathrm{z}_0), v(t; \mathrm{z}_0)\big)$ of (3.3) with $z(0) = \mathrm{z}_0$, which is defined on $\mathbb{R}$. Hence the Poincaré map

$$\phi : \mathcal{S} \to \mathcal{S}, \qquad \phi(\mathrm{z}_0) := z(T; \mathrm{z}_0)$$

is defined and it is continuous on $\mathcal{S}$. By the Liouville theorem it follows that ϕ is an area—preserving homeomorpism of the strip $\mathcal{S}$ onto itself. Clearly, all of these properties of ϕ hold true for any of the maps

$$\phi^k : \mathcal{S} \to \mathcal{S}, \quad \phi^k(\mathrm{z}_0) := z(kT; \mathrm{z}_0), \quad k \in \mathbb{Z}.$$

In particular, we note that z_0 is the initial point of a mT-periodic solutions $z(\cdot)$ of (3.3), with $m \in \mathbb{N}$, if and only if z_0 is a fixed point of the mth iterate ϕ^m of the Poincaré map ϕ.

A consequence of Proposition 3.2 which is crucial for the next application of the Poincaré–Birkhoff theorem is given by

$$\phi^{-m}((c,0)) = (\phi^m)^{-1}((c,0)) \in \mathcal{R}_m. \tag{3.10}$$

Finally, we set

$$\Psi_m(\mathrm{z}_0) := \frac{\psi_m(u(\cdot;\mathrm{z}_0))}{2\pi} \quad \text{for } \mathrm{z}_0 \in \mathcal{S} \setminus \mathcal{R}_m$$

and observe that the map

$$\Psi_m : \mathrm{dom}\Psi_m \supset \mathcal{S} \setminus \mathcal{R}_m \to \mathbb{R}$$

is continuous. Notice that if $u(\cdot;\mathrm{z}_0)$ is mT-periodic, then

$$\#_m u = 2\Psi_m(\mathrm{z}_0), \tag{3.11}$$

where $\#_m u$ is the number of zeros of $u(t) - c$ in the interval $[0, mT)$. In this case, the simplicity of the zeros of $u(t) - c$ implies that $\#_m u$ is always an even number.

With the above notation, we can prove the next result.

PROPOSITION 3.3. *Assume* (i$_1$), (i$_2$). *Then*

$$\Psi_m(\mathrm{z}_0) \to +\infty \quad \textit{as } \mathrm{z}_0 \to \partial\mathcal{S}.$$

Proof. In order to simplify the notation in the proof, we choose $c = (a+b)/2$.

From assumption (i$_2$) and (3.5), for any constant $R > 0$ we can find two numbers, d_R^- and d_R^+, with

$$A < d_R^- < a < c < b < d_R^+ < B$$

such that

$$g(x)(x-c) \geq 3R(x-c)^2 \quad \text{for all} \quad x \in (A, d_R^-] \cup [d_R^+, B).$$

Moreover, we can assume without loss of generality, by taking d_R^- smaller and d_R^+ larger, if necessary, that

$$\mathcal{R}_m \subset (d_R^-, d_R^+) \times \mathbb{R}.$$

Thus, for $x \in (A, d_R^-] \cup [d_R^+, B)$, and by choosing $R \geq \frac{1}{2}(2||E||_\infty/b-a)^2$, we have

$$\begin{aligned} g(x)(x-c) + y^2 + E(t)y &\geq 3R(x-c)^2 + \frac{1}{2}y^2 - \frac{1}{2}||E||_\infty^2 \\ &\geq 2R(x-c)^2 + \frac{1}{2}y^2 + R\left(\frac{b-a}{2}\right)^2 - \frac{1}{2}||E||_\infty^2 \\ &\geq 2R(x-c)^2 + \frac{1}{2}y^2. \end{aligned}$$

Now let us take $\gamma_R > 0$ so that

$$g(x)(x-c) \geq 3R(x-c)^2 - \gamma_R \quad \text{for all} \quad x \in (A,B).$$

Thus, for every $x \in [d_R^-, d_R^+]$, we obtain

$$g(x)(x-c) + y^2 + E(t)y \geq 2R(x-c)^2 + \tfrac{1}{2}y^2,$$

provided that $|y| \geq ||E||_\infty + (||E||_\infty^2 + 2\gamma_R)^{\frac{1}{2}} := D_R$.

In conclusion from this last argument, from Proposition 3.2, and from (3.8) we have that for all $m \in \mathbb{N}$, for all $R > 0$, there is a compact set $\mathcal{W}(R,m) \supset [d_R^-, d_R^+] \times [D_R^-, D_R^+]$ such that

$$\mathcal{R}_m \subset \mathcal{W}(R,m) \subset \mathcal{S}$$

and if $z(0) \in \mathcal{S} \setminus \mathcal{W}(R,m)$, then $z(t) \notin [d_R^-, d_R^+] \times [D_R^-, D_R^+]$ for $t \in [-mT, mT]$, and

$$-\theta'(t) \geq \frac{2R(u(t)-c)^2 + \frac{1}{2}v(t)^2}{(u(t)-c)^2 + v(t)^2}.$$

Thus, if $z(0) \notin \mathcal{W}(R,m)$ and $t \in [-mT, mT]$, we obtain

$$\frac{-\theta'(t)}{2R\cos^2\theta(t) + \frac{1}{2}\sin^2\theta(t)} \geq 1,$$

that implies

$$\int_{\theta(mT)}^{\theta(0)} \frac{d\theta}{2R\cos^2\theta + \frac{1}{2}\sin^2\theta} \geq mT$$

(see [3] for analogous computations). Using the fact that

$$\int_0^{2\pi} \frac{d\theta}{2R\cos^2\theta + \frac{1}{2}\sin^2\theta} = \frac{2\pi}{\sqrt{R}}$$

and recalling that

$$\frac{\theta(0) - \theta(mT)}{2\pi} \leq k+1,$$

where k denotes here the integer part of $\Psi_m(z(0))$, we obtain

$$k + 1 \geq mT\sqrt{R}/2\pi$$

and hence

$$\Psi_m(z(0)) \geq \left(mT\sqrt{R}/2\pi\right) - 1.$$

Letting R go to $+\infty$, we end the proof of the proposition. □

At this point we have all the tools to prove Theorem 3.1.

Proof of Theorem 3.1. We define the function

$$W : \mathcal{S} \to \mathbb{R}, \qquad W(x,y) := G(x) + \tfrac{1}{2}y^2$$

and observe that from (i$_1$) and (3.4) it follows that there is a constant $L_m > 0$ such that, for each $L \geq L_m$, the set $W^{-1}(L)$ is a simple closed curve which is star-shaped with respect to the point $(c, 0)$ and it is contained in $\mathcal{S} \setminus \mathcal{R}_m$. Then, for any r_1, r_2 with $L_m \leq r_1 < r_2$, we can consider the annulus

$$\mathcal{A} = \mathcal{A}(r_1, r_2) := \{(x, y) \in \mathcal{S} \mid r_1 \leq W(x, y) \leq r_2\} = W^{-1}\big([r_1, r_2]\big)$$

and the inner disc

$$\mathcal{D} = \mathcal{D}(r_1) := \{(x, y) \in \mathcal{S} \mid W(x, y) < r_1\} = W^{-1}\big((-\infty, r_1)\big).$$

By the choice of r_1 and r_2 we have that the boundary of $\mathcal{A}$ is the union of two simple closed curves $\partial^-\mathcal{A}$ and $\partial^+\mathcal{A}$, named respectively *the inner boundary* and *the outer boundary* of $\mathcal{A}$, such that

$$\partial\mathcal{D} = \partial^-\mathcal{A} = W^{-1}(r_1)$$

is star-shaped around $(c, 0)$. Moreover, as $\mathcal{D} \supset \mathcal{R}_m$, from (3.10) we obtain

$$(c, 0) \in \phi^m(\mathcal{D}).$$

Now we argue as follows.

At first we fix any constant $r = r(m) \geq L_m$ and, using the continuity of Ψ_m, define

$$\nu_m^* := \operatorname{int}\left[\max\{\Psi_m(\mathrm{z}_0) \mid \mathrm{z}_0 \in \partial\mathcal{D}(r)\}\right], \tag{3.12}$$

where $\operatorname{int}[\xi]$ denotes the integer part of the number $\xi \in \mathbb{R}$.

Second, we choose any number $p \in \mathbb{N}$, with p prime with m and

$$p > \nu_m^*. \tag{3.13}$$

Then, using Proposition 3.3 and the continuity of Ψ_m, we can find another constant $R = R(m, p)$ with $R > r$ such that

$$\min\{\Psi_m(\mathrm{z}_0) \mid \mathrm{z}_0 \in W^{-1}(R)\} > p. \tag{3.14}$$

Now we observe that (3.12), (3.13), and (3.14) imply that on the boundaries of the annulus

$$\mathcal{A} = \mathcal{A}(r, R),$$

the twist condition

$$\Psi_m(\mathrm{z}_0) < p \quad \text{for } \mathrm{z}_0 \in \partial^-\mathcal{A}, \qquad \Psi_m(\mathrm{z}_0) > p \quad \text{for } \mathrm{z}_0 \in \partial^+\mathcal{A}$$

is satisfied.

Thus we have met all the conditions in order to apply W. Ding's generalization of the Poincaré–Birkhoff fixed point theorem [10] and hence we can conclude that the map ϕ^m has a fixed point, say $\mathrm{z}_{m,p}^*$, belonging to the annulus $\mathcal{A}$. Furthermore, we also obtain

$$\Psi_m\big(\mathrm{z}_{m,p}^*\big) = p. \tag{3.15}$$

The continuity of Ψ_m implies that $\Psi_m\big(\mathrm{z}_0\big)$ is bounded for z_0 belonging to a compact subset of $\mathcal{S} \setminus \mathcal{R}_m$, thus, as

$$\Psi_m\big(\mathrm{z}_{m,p}^*\big) \to +\infty$$

for $p \to +\infty$, we have

$$\mathrm{z}^*_{m,p} \to \partial\mathcal{S} \quad \text{as} \quad p \to +\infty.$$

This last property, in connection with Proposition 3.2, finally implies that

$$\left(u^*_{m,p}(t), \frac{d}{dt}u^*_{m,p}(t)\right) \to \partial\mathcal{S} \quad \text{as} \quad p \to +\infty,$$

uniformly with respect to $t \in [0, mT]$, where $(u^*, v^*) = z^*$ is the solution to (3.3) starting at $\mathrm{z}^* = \mathrm{z}^*_{m,p}$ at the time $t = 0$.

Our last goal now is to prove that the solution we find does have minimal period mT. To this aim it is sufficient to prove that

$$\phi^k\left(\mathrm{z}^*_{m,p}\right) \neq \mathrm{z}^*_{m,p} \quad \text{for each} \quad 1 \leq k \leq m-1,$$

holds.

Assume by contradiction (cf. [7, §6]) that there is some k with $1 \leq k < m$ such that $\mathrm{z}^* = \mathrm{z}^*_{m,p}$ is a fixed point of ϕ^k. Then, by (3.9), (3.11), and the definition of Ψ_m, we obtain that

$$\exists\, \ell \in \mathbb{N},\ \ell < p: \quad \Psi_k(\mathrm{z}^*) = \ell.$$

Observing now that

$$(\phi^m)^k(\mathrm{z}^*) = (\phi^k)^m(\mathrm{z}^*) = \mathrm{z}^*,$$

we have

$$pk = \Psi_{m\cdot k}(\mathrm{z}^*) = \ell m$$

which yields

$$\frac{m}{p} = \frac{k}{\ell},$$

a contradiction with the assumption that p is prime with m.

In conclusion, we observe that if the local uniqueness of the solutions for the Cauchy problems associated to (3.3) is not guaranteed, we have to repeat the above argument for a sequence of approximating equations of the form

$$u' = v + E(t), \qquad v' = -g_n(u),$$

where $g_n : (A, B) \to \mathbb{R}$ is smooth and $g_n \to g$ uniformly on compact sets. It is possible to check that, for n sufficiently large, all the fixed points of the iterates of the Poincaré operator of the approximating equations belong to the same annulus $\mathcal{A}$ and the rotation number Ψ_m of all these fixed points is the same and equal to p. Hence we can pass to the limit for a subsequence and get a fixed point of ϕ^m (which now could be a multivalued function). For the missing technical details concerning such an approximation approach, we refer the reader to [7, §6].

At this step, all the assertions in Theorem 3.1 are justified and the proof is now complete. □

4. Examples. In this section we consider two examples for the applicability of our main results.

First, we examine the case with one singularity at a point $A \in \mathbb{R}$.

Example 4.1. Let an electric charge Q be placed at the fixed point $A \in \mathbb{R}$ and suppose that $y > A$ (or $y < A$) denotes the position of an electric charge q, having the same sign of Q, which lives in a one-dimensional space.

The Coulombian force h acting on q at time t is given by

$$h(y(t)) = \kappa q Q \frac{y(t) - A}{|y(t) - A|^3} := \kappa_0 \frac{y(t) - A}{|y(t) - A|^3},$$

where κ is a suitable constant and $\kappa_0 = \kappa q Q > 0$.

Let $e : \mathbb{R} \to \mathbb{R}$ be a T-periodic external forcing term acting on the system. We assume that $e \in L^1_{\mathrm{loc}}$ and denote by $\bar{e} = \frac{1}{T} \int_0^T e(s)\, ds$ the mean value of the function e.

Then the Newton law for the dynamics of the charge q yields the following differential equation (unitary mass is assumed).

$$y''(t) = h(y(t)) + e(t). \tag{4.1}$$

Now, from Theorem 2.1, we have the following proposition.

PROPOSITION 4.1. *Equation* (4.1) *has periodic solutions if and only if* $\bar{e} \neq 0$. *If* $\bar{e} < 0$ (*respectively,* $\bar{e} > 0$), *all periodic solutions of* (4.1) *lie in* $(A, +\infty)$ (*respectively,* $(-\infty, A)$), *and besides having at least one T-periodic solution,* (4.1) *also has subharmonic solutions with minimal period mT, for every sufficiently large integer m.*

Proof. We rewrite (4.1) as

$$u''(t) + g(u(t)) = e(t), \tag{4.2}$$

with $u(t) = y(t) - A$ and $g(x) := -h(x + A)$. Let G be a primitive of g, e.g.,

$$G(x) = \frac{\kappa_0}{|x|}.$$

Note that

$$\lim_{|x| \to \infty} g(x) = 0, \qquad \lim_{|x| \to \infty} \frac{G(x)}{x^2} = 0 \tag{4.3}$$

and that

$$\lim_{x \to 0} g(x) \mathrm{sgn}(-x) = \lim_{x \to 0} G(x) = +\infty. \tag{4.4}$$

At first we claim that $\bar{e} \neq 0$ is a necessary condition. This follows from [17]. Namely, assume for instance that (4.2) has a τ-periodic solution $\tilde{u}(t)$, with $\tilde{u}(t) > 0$ for all $t \in \mathbb{R}$. Integrating both sides of (4.2) on the interval $[0, \tau]$, we obtain

$$\frac{1}{\tau} \int_0^\tau g(\tilde{u}(t))\, dt = \frac{1}{\tau} \int_0^\tau e(t)\, dt = \bar{e}.$$

Since $g(x) < 0$ for all $x > 0$, we obtain $\bar{e} < 0$. Analogously, if (4.2) has a negative solution, $\bar{e}$ has to be positive. Thus the claim is achieved.

Now, we suppose $\bar{e} < 0$, According to (4.3) and (4.4) all the assumptions of Theorem 2.2 are satisfied and we conclude with the result. If $\bar{e} > 0$, we can reduce by a change of variables to the previous case (see Remark 2.2). □

Next we present an example with two singularities.

Example 4.2. Let Q_1 and Q_2 be two electric charges placed at the fixed points $A \in \mathbb{R}$ and $B \in \mathbb{R}$ with $A < B$. Suppose that u, with $A < u < B$ denotes the position of an electric charge q where we assume that Q_1, Q_2 and q have all the same sign.

Now the Coulombian force l acting on q at the time t takes the form

$$\begin{aligned} l(u(t)) &= \kappa\frac{qQ_1}{(u(t)-A)^2} - \kappa\frac{qQ_2}{(u(t)-B)^2} \\ &= \frac{\kappa_1}{(u(t)-A)^2} - \frac{\kappa_2}{(u(t)-B)^2}, \end{aligned}$$

where $\kappa_i = \kappa q Q_i > 0$ for $i = 1, 2$.

Let $e : \mathbb{R} \to \mathbb{R}$ be a T-periodic forcing term as above and consider the differential equation

$$u''(t) = l(u(t)) + e(t). \tag{4.5}$$

We assume also that $\int_0^t (e(s) - \bar{e})\, ds$ is not constant. Then from Theorem 3.1 we have the following proposition.

PROPOSITION 4.2. *For every $m \geq 1$, equation* (4.5) *has infinitely many mT-periodic solutions lying in (A, B), all having minimal period mT.*

Proof. We write (4.5) as

$$u''(t) + g(u(t)) = e(t). \tag{4.6}$$

with $g(x) := -l(x)$. Note that

$$\lim_{x\to A^+} g(x) = -\infty, \qquad \lim_{x\to B^-} g(x) = +\infty$$

and

$$\lim_{x\to A^+} G(x) = \lim_{x\to B^-} G(x) = +\infty,$$

where, for a fixed $c \in (A, B)$,

$$G(x) = \int_c^x g(s)\, ds = \frac{\kappa_1}{x-A} + \frac{\kappa_2}{B-x} - \frac{\kappa_1}{c-A} - \frac{\kappa_2}{B-c}.$$

Since $e(t)$ is T-periodic and nonconstant, its minimal period equals to T/γ, for some $\gamma \in \mathbb{N}$. We apply now Theorem 3.1 and have that for every $m \geq 1$ equation (4.6) has at least infinitely many periodic solutions having $mT = m\gamma(T/\gamma)$ as minimal period. □

Remark 4.1. Since the nonlinearities in the above examples are locally Lipschitz continuous, arguing as in [21] we could claim that for any subharmonic solution we found there is a second one with the same minimal period and the same number of zeros, which is not a shift of the previous one.

REFERENCES

[1] A. AMBROSETTI AND V. COTI ZELATI, *Solutions with minimal period for Hamiltonian systems in a potential well*, Ann. Inst. H. Poincaré Anal. Non Lineaire, 4 (1987), pp. 275–296.

[2] V. BENCI, *Normal modes of a Lagrangian system constrained in a potential well*, Ann. Inst. H. Poincaré Anal. Non Lineaire, 1 (1984), pp. 379–400.

[3] A. Capietto, J. Mawhin, and F. Zanolin, *A continuation approach to superlinear periodic boundary value problems*, J. Differential Equations, 88 (1990), pp. 347–395.

[4] V. Coti Zelati, S. Li, and S. Wu, *Periodic solutions of a class of singular nonautonomous second order systems in potential well*, SISSA 1991/M, preprint.

[5] M. A. Del Pino and R. F. Manásevich, *Infinitely many T-periodic solutions for a problem arising in nonlinear elasticity*, J. Differential Equations, to appear.

[6] M. A. Del Pino, R. F. Manásevich, and A. E. Murua, *On the number of 2π-periodic solutions for $u'' + g(u) = s(1 + h(t))$ using the Poincaré–Birkhoff theorem*, J. Differential Equations, 95 (1992), pp. 240–258.

[7] T. Ding and F. Zanolin, *Periodic solutions of Duffing's equation with superquadratic potential*, J. Differential Equations, 97 (1992), pp. 328–378.

[8] T. Ding and F. Zanolin, *Subharmonic solutions of second order nonlinear equations: a time-map approach*, Nonlinear Analysis, TMA, 20 (1993), pp. 509–532.

[9] W. Ding, *Fixed points of twist mappings and periodic solutions of ordinary differential equations*, Acta Math. Sinica, 25 (1982), pp. 227–235. (In Chinese.)

[10] ———, *A generalization of the Poincaré–Birkhoff theorem*, Proc. Amer. Math. Soc., 88 (1983), pp. 341–346.

[11] A. Fonda and A. C. Lazer, *Subharmonic solutions of conservative systems with non-convex potential*, Proc. Amer. Math. Soc., 115 (1992), pp. 183–190.

[12] A. Fonda and M. Ramos, *Large amplitude subharmonic oscillations for scalar second order differential equations with asymmetric nonlinearities*, J. Differential Equations, to appear.

[13] A. Fonda, M. Ramos, and M. Willem, *Subharmonic solutions for second order differential equations*, Topological Methods in Nonlinear Anal., 1 (1993), pp. 41–57.

[14] P. Habets and L. Sanchez, *Periodic solutions of some Lienard equations with singularities*, Proc. Amer. Math. Soc., 109 (1989), pp. 1035–1046.

[15] J.K. Hale, *Ordinary Differential Equations*, R. E. Krieger Publishing, Huntington, NY, 1980.

[16] J.P. La Salle and S. Lefschetz, *Stability by Liapunov's Direct Method with Applications*, Academic Press, New York, 1963.

[17] A.C. Lazer and S. Solimini, *On periodic solutions of nonlinear differential equations with singularities*, Proc. Amer. Math. Soc., 88 (1987), pp. 109–114.

[18] J. Mawhin, *Topological degree and boundary value problems for nonlinear differential equations*, CIME Lectures, Montecatini, June 1991, to appear.

[19] G. R. Morris, *A differential equation for undamped forced oscillations* III, Proc. Cambridge Phil. Soc., 61 (1965), pp. 133–155.

[20] ———, *An infinite class of periodic solutions of $x'' + 2x^3 = p(t)$*, Proc. Cambridge Phil. Soc., 61 (1965), pp. 157–164.

[21] W. D. Neumann, *Generalizations of the Poincaré–Birkhoff fixed point theorem*, Bull. Austral. Math. Soc., 17 (1977), pp. 375–389.

[22] P. Omari, G. Villari, and F. Zanolin, *Periodic solutions of the Liénard equation with one-sided growth restrictions*, J. Differential Equations, 67 (1987), pp. 278–293.

[23] M. Struwe, *Multiple solutions of anticoercive boundary value problems for a class of ordinary differential equations of second order*, J. Differential Equations, 37 (1980), pp. 285–295.

SIAM J. MATH. ANAL.
Vol. 24, No. 5, pp. 1312–1330, September 1993

011

BOUNDEDNESS OF PRIME PERIODS OF STABLE CYCLES AND CONVERGENCE TO FIXED POINTS IN DISCRETE MONOTONE DYNAMICAL SYSTEMS*

PETER HESS† AND PETER POLÁČIK‡

Abstract. In this paper the boundedness of minimal periods of linearly stable cycles for discrete, strongly order-preserving semigroups $(F_0^n)_{n\in\mathbb{N}}$ in bounded subsets of an ordered Banach space is proved. It is further shown that this bound is not increased by small perturbations of F_0. Of particular interest is the case where the only linearly stable cycles of F_0 are fixed points. Employing a recent result of Poláčik and Tereščák, the typical convergence of relatively compact orbits and for perturbed systems then follow. The results are applied to classes of time-periodic reaction-diffusion equations and give typical convergence to periodic solutions.

Key words. discrete monotone dynamical systems, linearly stable cycles, Lyapunov exponents, center manifold theorem, Krein–Rutman theorem, periodic reaction-diffusion equations, generic convergence

AMS subject classifications. primary 58F39, 35B40, 35K57; secondary 58F22, 34C35, 54H20, 35B10

1. Introduction. In the recent paper [PT1], Poláčik and Tereščák address the question of the typical asymptotic behavior of orbits in strongly monotone discrete-time dynamical systems. In contrast to the continuous-time case, where a rather complete description of the typical behavior of trajectories is available (see [Hi1], [Hi3], [Ma], [ST1], [ST2], [Sm], [T2], [P1], [P2], [Mi1], or the introduction of [PT1] for a brief discussion of these results), [PT1] seems to be the first paper that provides an answer to this question for a rather general class of smooth ($C^{1,\alpha}$) discrete systems. The authors prove that a typical trajectory (i.e., a trajectory emanating from a residual set of initial conditions) of such a discrete system converges to a linearly stable cycle. (Here cycle refers to the orbit of a periodic point, and linearly stable means that the spectrum of the linearization of an appropriate iterate of the mapping lies in the closed unit disc in $\mathbb{C}$; see §2 for the precise definition.) This result applies to dynamical systems generated by the period map of many types of periodically forced differential equations, including scalar parabolic equations and cooperative systems of parabolic equations on bounded domains, and to cooperative systems of ODEs.

It is known that, unless severe restrictions (like those in [AH1], [AH2], [AHM], [Hs1], [Hs2], [T1], [T5]) are imposed, the result of [PT1] cannot be improved as to assert typical convergence to fixed points rather than to cycles. Counterexamples can be found in [DH], [T3], [T4]. However, an improvement is possible. We show here that one can establish typical convergence to a fixed point for a certain iterate F_0^m of the mapping F_0 in question. This is true at least for any bounded set in the state space (with m depending on this set). Specifically, we prove that for any bounded set B,

*Received by the editors September 25, 1991; accepted for publication (in revised form) December 1, 1992.

†Mathematisches Institut, Universität Zürich, 8001 Zürich, Switzerland, until his death in November 1992.

‡Institute of Applied Mathematics, Comenius University, Mlynská dolina, 842 15 Bratislava, Czechoslovakia. This work was performed while the author was visiting the Mathematics Institute of the University of Zürich. His stay was supported in part by the Swiss National Science Foundation.

the set of minimal periods of linearly stable cycles contained in B is bounded. This, in conjunction with the result of [PT1], implies the above improvement.

We obtain boundedness of the minimal periods of linearly stable cycles (below such periods are called stable periods) as a by-product in proving a certain perturbation result. Roughly speaking, this result says that for any bounded domain B, the maximal stable period of F_0 in B is not increased by a small perturbation of F_0. More precisely, if $F_\varepsilon, \varepsilon \in \mathbb{R}$, is a family of mappings, satisfying the hypotheses of §2, then the maximal stable period of F_0 (in B) is not less than the maximal stable period of F_ε for any ε sufficiently close to zero.

This result is of particular interest in case F_0 does not have stable periods greater than 1 (i.e., all linearly stable cycles are fixed points of F_0). It is well known (see [Hs2, Prop. 9.4]) that this situation occurs when F_0 is a time-t map of a strongly monotone semiflow. Thus our perturbation result can be applied to small time-periodic perturbations of autonomous equations that generate a strongly monotone semiflow (see §5 for an example). Moreover, there are several different classes of discrete dynamical systems that are known not to have higher stable periods. We discuss two such classes and their perturbations in §5.

Our main results and some immediate consequences are stated in §2. Section 3 gives a preparation for the proofs of the main results which are then carried out in §4.

2. Statement of the main results.

In the whole paper X is a strongly ordered Banach space with norm $\|\cdot\|$ and order cone X_+. By an order cone we mean a closed convex cone such that $X_+ \cap (-X_+) = \{0\}$, and X being strongly ordered requires $\operatorname{int} X_+$, the interior of X_+, to be nonempty. For $x, y \in X$ we write

$$\begin{aligned} x &\le y && \text{if } y - x \in X_+, \\ x &< y && \text{if } x \le y \quad \text{and} \quad x \ne y, \\ x &\ll y && \text{if } y - x \in \operatorname{int} X_+. \end{aligned}$$

The reversed signs are used in the usual way. A mapping $h : X \to X$ is called monotone if $x \le y$ implies $h(x) \le h(y)$.

We can now formulate our hypotheses.

(H) For each $\varepsilon \in J$, where $J = [-\delta, \delta]$ is an interval in $\mathbb{R}, F_\varepsilon : X \to X$ is a compact monotone mapping such that the following properties hold:

(a) $F : (\varepsilon, x) \mapsto F_\varepsilon(x) : J \times X \to X$ is of class C^1;

(b) $F_0 : X \to X$ is of class $C^{1,\theta}$ (C^1 with locally θ-Hölder continuous derivative), $\theta \in (0, 1]$, and it is one-to-one. For any $x \in X$ the differential $d_x F_0 = F_0'(x)$ is a strongly positive operator (i.e., $v > 0$ implies $d_x F_0 v \gg 0$).

Notice that compactness of each F_ε (by which we mean that each bounded set is mapped onto a relatively compact set), in conjunction with (H)(a), implies that $F : J \times X \to X$ is a compact map.

Below, several results of [PT1] are applied to F_0. To justify this, we make a few comments on the hypotheses imposed on F_0.

The standing hypotheses of [PT1], see p. 340, are identical to our (H)(b) without the assumption of F_0 being one-to-one. There is another hypothesis, existence of continuous separation along any compact set invariant under F_0 (see [PT1, p. 344]), which (or whose weaker form) is used at various places in [PT1]. However, as was shown in [PT2], this hypothesis follows from (H)(b)—the injectivity assumption is essential for this. We can thus use the results of [PT1] freely.

For the formulation of our results we need some more definitions. Let $h : X \to X$ be a C^1-map (below we always consider $h = F_0$ or $h = F_\varepsilon$). A point $x \in X$ is a *periodic point of* h if $h^k(x) = x$ for some natural number k, which is then a *period* for x. If it is the minimal period, i.e., if $h^m(x) \neq x$ for $m = 1, \dots, k-1$, we call x *k-periodic.* A k-periodic point x is said to be *linearly stable* if the spectrum of the operator $d_x h^k$ is contained in the closed unit disc in $\mathbb{C}$. Note that "linearly stable" refers only to the location of the spectrum and such a point may be unstable for the dynamical system defined by h. A more precise expression would be "linearly stable or linearly neutrally stable." For brevity, we use the former expression throughout.

Let $B \subset X$. We say that k is a *stable period* for the restriction $h|\hat{B}$ if there is a linearly stable k-periodic point x of h such that the orbit $O(x,h) := \{h^n(x) : n = 0, 1, \dots\} = \{h^n(x) : n = 0, 1, \dots, k-1\}$ is contained in B. If $B = X$ we simply say that k is a stable period for h. Similarly as for the orbits $O(x,h)$, we always indicate to which map the dynamical notions refer. We, e.g., write

$$\omega(x,h) := \{y \in X : \text{ there is a sequence } n_j \to +\infty \text{ such that } h^{n_j}(x) \to y\}$$

for the *ω-limit set* of x with respect to h.

Our main theorems can now be formulated as follows.

THEOREM 1. *Let* (H)(b) *hold. Let $B \subset X$ be a closed bounded set. Then there exists a constant m such that the set of stable periods of $F_{0|B}$ is bounded above by m.*

THEOREM 2. *Let* (H) *hold. Let $B \subset X$ be a closed bounded set. Suppose that all the stable periods of $F_{0|B}$ are bounded above by m. Then there exists a $\delta_0 > 0$ such that for any $\varepsilon \in [-\delta_0, \delta_0]$ all the stable periods of $F_{\varepsilon|B}$ are bounded by m.*

The following result is an important particular case of Theorem 2.

COROLLARY 3. *Let* (H) *hold. Suppose that all linearly stable periodic points of F_0 are fixed points of F_0 (i.e., there are no stable periods for F_0 larger than* 1). *Then, given any closed bounded set B, there exists a $\delta_0 > 0$ such that for any $\varepsilon \in [-\delta_0, \delta_0]$ all linearly stable periodic points of $F_{\varepsilon|B}$ are fixed points of F_ε.*

Of course if it is known a priori that all periodic points of F_ε lie in a fixed bounded set, then the statement of this corollary holds for $B = X$.

Theorem 1 allows for an improvement of the result on generic convergence to cycles proved by Poláčik and Tereščák [PT1, Thm. 5.1]. This result says that in any open bounded set $\mathcal{G}$ positively invariant under F_0, there is an open dense subset of points whose ω-limit sets are linearly stable cycles. This and Theorem 1 (applied to $\overline{\mathcal{G}}$) immediately give the following.

COROLLARY 4. *Let* (H)(b) *hold. Let $\mathcal{G}$ be an open bounded set positively invariant under F_0 (i.e., $F_0\mathcal{G} \subset \mathcal{G}$). Then there is an m such that the set*

$$\{x \in \mathcal{G} : \omega(x, F_0) \textit{ is a cycle of period at most } m\}$$

contains an open and dense subset of $\mathcal{G}$.

Here cycle of period k means the orbit of a k-periodic point.

Finally, we formulate a theorem which implies that in the class of mappings satisfying (H)(b) generic convergence is an open property. A point $x \in X$ is said to be *convergent* for h if $h^n(x)$ converges, as $n \to +\infty$, to a $z \in X$. Another way to say this is that the orbit $O(x,h)$ is relatively compact and $\omega(x,h)$ consists of a single point z, which is necessarily a fixed point of h.

COROLLARY 5. *Let* (H) *hold. In addition suppose that* (H)(b) *holds with F_0 replaced by F_ε for any $\varepsilon \in [-\delta, \delta]$. Let $\mathcal{G}$ be an open bounded set such that for each $\varepsilon \in [-\delta, \delta]$ all the orbits $O(x, F_\varepsilon), x \in \mathcal{G}$, lie in a ball D independent of ε. Finally,* *suppose that all linearly stable periodic points of $F_0|_{\overline{D}}$ are fixed points. Then there exists a $\delta_0 > 0$ such that for any $\varepsilon \in [-\delta_0, \delta_0]$ the set*

$$\{x \in \mathcal{G} : x \textit{ is convergent for } F_\varepsilon\}$$

contains an open and dense subset of $\mathcal{G}$.

This result follows from [PT1, Thm. 5.1] applied to F_ε and from the fact that all stable periods of F_ε for ε near zero equal 1 (cf. Theorem 2).

An important example of F_0 which satisfies the requirement that all linearly stable periodic points are fixed points is the time-one map of a strongly monotone (continuous-time) semiflow [Hs2, Prop. 9.4]). Thus Corollary 5 can be applied to small nonautonomous time-periodic perturbations of an autonomous differential equation generating a strongly monotone semiflow. A different type of application we present in §5 establishes generic convergence for equations which are perturbations of some equations taken in a special class where generic convergence is known to hold. Such perturbations are allowed to fall outside this special class.

The proofs of Theorems 1, 2 consist of the following two steps: (I) localization, (II) local bifurcation.

In the first step, the localization, we prove that any sequence p_n of linearly stable periodic points of $F_{\varepsilon_n|B}$ with $\varepsilon_n \to 0$ (in particular one can take $\varepsilon_n = 0$) contains a subsequence p_{n_j} such that

$$\operatorname{dist}(O(p_{n_j}, F_{\varepsilon_{n_j}}), \qquad O(z, F_0)) \to 0 \quad \text{as } j \to +\infty.$$

Here z is a linearly stable periodic point for F_0 and dist is as in (4.3) below. So for this subsequence, the orbit of p_{n_j} is as close to the orbit $O(z, F_0)$ as we wish if j is large enough.

In the second step, we then consider such a local situation and prove that for large j the minimal period of p_{n_j} cannot be larger than that of z. This will conclude the proof of Theorem 2 and also the proof of Theorem 1 because this shows that there cannot be a sequence of stable periods for $F_{0|B}$ converging to $+\infty$.

In the first step investigation of a "limit set" of $O(p_\varepsilon, F_\varepsilon)$ as $\varepsilon \to 0$ and Lyapunov exponents for points in this limit set are the crucial ingredients.

The second step is just a combination of the center manifold theorem and the Krein–Rutman theorem.

We carry out the two steps after the following preliminary section.

3. Preliminaries. In this section we fix notation and prove or recall some basic assertions needed in the sequel. Let $h : X \to X$ be a C^1-map. If $x \in X$ and the orbit $O(x,h)$ is relatively compact, then the ω-limit set $\omega(x,h)$ is a nonempty, compact set invariant under h. If h has strongly positive differentials, i.e., $d_x hv = h'(x)v \gg 0$ for any $x \in X, v > 0$ (which is the case for $h = F_0$), then $\omega(x,h)$ is an *unordered set.* More precisely, there are no two points $z, y \in \omega(x,h)$ with $z < y$. This is a general property valid for any *strongly monotone* mapping, that is a mapping which takes related points $x < y$ onto strongly relaxed points $h(x) \ll h(y)$ (see [T1], [T5], [Hi2]). A mapping h with strongly positive differentials is strongly monotone. To see this one just observes that for any $x < y$ the curve $\{h(x + s(y-x)) : \ s \in (-1,2)\}$ is at any of its points z tangent to $d_z h(y-x) \gg 0$, hence any two distinct points on this curve are related by $\ll$.

A consequence of this unorderedness property of the ω-limit sets is that if x is a periodic point for a strongly monotone map h, then $O(x,h)$, which equals $\omega(x,h)$, does not contain two related points. This property of periodic points remains valid under the weaker assumption that h is merely monotone (not necessarily strongly). This can be proved as follows; cf. [Hi4, Prop. 0(a)]. Suppose that x is a k-periodic point of h and that $O(x,h)$ contains two related points, i.e.,

$$h^{m+r}x < h^m x$$

for some $m \in \{0,1,\dots\}$ and $r \in \{1,\dots,k-1\}$. Then by [Hi2, Prop. 6.1], $\omega(h^m x, h)$ is a cycle with the minimal period at most r. This is a contradiction because $\omega(h^m x, h) = O(x,h)$ and r is less than k, the minimal period of x.

Unorderedness of cycles will be later used for $h = F_\varepsilon$. We will also use the fact that if h is C^1 and monotone, then $d_x h$ is a positive linear operator for any x. Indeed, for any $v \geq 0$ one has

$$d_x h v = \lim_{s \to 0+} \frac{h(x+sv) - h(x)}{s} \geq 0$$

by monotonicity of h and closedness of the relation $\geq$. This of course implies that $d_x h^n$ is positive for any $x \in X$ and $n = 0,1,2,\dots$.

For any $x \in X, v \in X$ we define

$$\lambda(x,v,h) = \limsup_{n \to +\infty} \frac{\log \|d_x h^n v\|}{n} \tag{3.1}$$

and

$$\lambda_1(x,h) = \sup_{\substack{v \in X \\ v \neq 0}} \lambda(x,v,h).$$

We call $\lambda_1(x,h)$ the first *Lyapunov exponent* of x (with respect to h). This is a slight abuse of language because in general one does not know if $\lambda_1(x,h) = \lambda(x,h,v)$ for some v (which is a Lyapunov exponent in the usual terminology). For $h = F_0$, the hypothesis (H)(b) implies that $\lambda_1(x,F_0) = \lambda(x,v,F_0)$ for any $v > 0$ (see [PT1, §3]; for a general background in Lyapunov exponents see [M1], [M2]).

If x is a k-periodic point of h and h is compact (as is the case for $h = F_\varepsilon$), then

$$\lambda_1(x,h) = \frac{1}{k} \log \operatorname{spr}(d_x h^k), \tag{3.2}$$

where $\operatorname{spr}(d_x h^k)$ is the spectral radius of $d_x h^k$. Moreover,

$$\lambda_1(x,h) = \lambda(x,h,v), \tag{3.3}$$

where v is an eigenvector (or the real part of an eigenvector) corresponding to any eigenvalue μ with $|\mu| = \operatorname{spr}(d_x h^k)$. These properties follow from the equality

$$d_x h^{mk+r} = d_x h^r d_x h^{mk} = (d_x h^r)(d_x h^k)^m \tag{3.4}$$

obtained by the chain rule and periodicity of x. Indeed, substituting (3.4) in (3.1) with $n = km + r, m \in \{0,1,\dots\}, r \in \{0,1,\dots,k-1\}$ and using the fact that r varies in a finite set, we obtain

$$\lambda(x,v,h) \leq \frac{1}{k} \log \operatorname{spr}(d_x h^k) \tag{3.5}$$

for any $v \neq 0$. Therefore, $\lambda_1(x,h)$ is not larger than this logarithm. On the other hand, if v is the real part of an eigenvector as above, then

$$\begin{aligned}\lambda(x,v,h) &\geq \limsup_{m\to\infty} \frac{\log \|(d_x h^k)^m v\|}{km} \\ &\geq \frac{1}{k}\log|\mu| = \frac{1}{k}\log \operatorname{spr}(d_x h^k).\end{aligned}$$

This and (3.5) imply (3.2) and (3.3).

By (3.2), a periodic point x of h is linearly stable if and only if

$$\lambda_1(x,h) \leq 0.$$

Below we consider only compact mappings $h = F_\varepsilon$, and we use the latter equivalent definition of linear stability.

We now introduce the notion of a *regular point* of the strongly monotone compact map F_0. We say that $x \in X$ is regular if

$$\lambda(x,v,F_0) = \lim_{n\to+\infty} \frac{\log \|d_x F_0^n v\|}{n}$$

for any $v > 0$, i.e., $\lambda_1(x,F_0) = \lambda(x,v,F_0)$ is the limit, not just the superior limit, for any $v > 0$.

We conclude this section by proving the following "continuity" property of the first Lyapunov exponent.

LEMMA 3.1. *Let* (H) *hold. Let $K \subset X$ be a compact set invariant under F_0. Suppose that*

$$\lambda_1(z,F_0) > 0 \quad \textit{for each } z \in K. \tag{3.6}$$

Then there exist a $\delta_1 > 0$ and a neighborhood U of K with the following property: If $\varepsilon \in [-\delta_1, \delta_1]$ and $y \in U$ is such that $O(y,F_\varepsilon) \subset U$, then $\lambda_1(y,F_\varepsilon) > 0$.

Proof. Fix a vector $w \gg 0$. By [PT1, Proof of Prop. 4.5] (see formula (4.2) of [PT1]), for any $z \in K$ there is a ν such that

$$d_z F_0^\nu w \gg 2w. \tag{3.7}$$

We now show that (3.7) implies the following claim.

There exist a $\delta_1 > 0$, a neighborhood U of K, and an integer-valued bounded function $y \mapsto \nu(y)$ defined on U such that

$$d_y F_\varepsilon^\nu w \gg 2w \text{ whenever } y \in U, \quad \varepsilon \in [-\delta_1,\delta_1], \text{ and } \quad \nu = \nu(y). \tag{3.8}$$

To prove the claim we first fix a $z \in K$. Let ν be as in (3.7). Since $(\varepsilon, y) \to d_y F_\varepsilon^\nu$ is continuous and (3.7) is an open relation, there is a $\delta(z) > 0$ and a neighborhood $U(z)$ of z such that

$$d_y F_\varepsilon^\nu w \gg 2w \quad \text{for any } y \in U(z), \quad \varepsilon \in [-\delta(z), \delta(z)].$$

Taking a finite cover of K by such neighborhoods and letting U be the union of the cover and δ_1 the minimum of the corresponding $\delta(z)'s$, we find a function $\nu(y)$ on U as desired.

Now we verify that U and δ_1 just constructed have the properties asserted in Lemma 3.1. Fix $\varepsilon \in [-\delta_1, \delta_1]$ and $y \in U$ such that $O(y, F_\varepsilon) \subset U$. Let

$$y_n := F_\varepsilon^n(y), \qquad n = 0, 1, \dots,$$

and define the sequence $n_k, k = 1, 2, \dots$, as follows:

$$\begin{aligned} n_1 &:= \nu(y), \\ n_{k+1} &:= n_k + \nu(y_{n_k}). \end{aligned}$$

This makes sense since $y_n \in U$ for any n. By (3.8) we have

$$d_y F_\varepsilon^{n_1} w \gg 2w. \tag{3.9}$$

Using the chain rule we further obtain

$$d_y F_\varepsilon^{n_{k+1}} w = d_{y_{n_k}} F_\varepsilon^{\nu_k} d_y F_\varepsilon^{n_k} w, \tag{3.10}$$

where $\nu_k = \nu(y_{n_k})$. Since each $d_{y_{n_k}} F_\varepsilon^{\nu_k}$ is positive, from (3.9), (3.10) we obtain by induction with respect to k that

$$d_y F_\varepsilon^{n_k} w \geq 2^k w. \tag{3.11}$$

Let m be an upper bound for $\nu(y)$, $y \in U$. Then

$$n_k = \nu(y) + \nu(y_{n_1}) + \cdots + \nu(y_{n_{k-1}}) \leq mk,$$

and hence

$$k \geq n_k/m.$$

This and (3.11) lead to

$$(e^{-\gamma})^{n_k} d_y F_\varepsilon^{n_k} w \geq w \gg 0, \tag{3.12}$$

where

$$\gamma = \log\left(2^{\frac{1}{m}}\right) > 0.$$

Inequality (3.12) implies that $(e^{-\gamma})^{n_k} \|d_y F_\varepsilon^{n_k} w\|$ stays bounded away from zero (otherwise, by closedness of the relation $\leq$, it would follow that $w = 0$). This implies

$$\lambda(y, w, F_\varepsilon) \geq \gamma > 0,$$

and consequently,

$$\lambda_1(y, F_\varepsilon) \geq \lambda(y, w, F_\varepsilon) > 0,$$

as claimed. □

4. Proofs of Theorems 1 and 2. The proofs have two common steps as outlined in §2. Throughout the section we assume that (H) holds.

I. Localization. Let $B \subset X$ be a closed bounded set. Let $\varepsilon_n, n = 1, 2, \dots$, be a sequence in $J = [-\delta, \delta]$ converging to zero. To minimize the number of indices we shall use the notation

$$F_n = F_{\varepsilon_n}.$$

Suppose that for $n = 1, 2, \ldots$, p_n is a k_n-periodic point of F_n such that

$$O(p_n, F_n) \subset B \tag{4.1}$$

and

$$\lambda_1(p_n, F_n) \leq 0 \quad \text{(i.e., } p_n \text{ is linearly stable).} \tag{4.2}$$

The aim of *localization* is to prove that there exists a linearly stable periodic point z of F_0 and a sequence $n_j \to +\infty$ such that $O(z, F_0) \subset B$ and

$$\operatorname{dist}(O(p_{n_j}, F_{n_j}), \qquad O(z, F_0)) \to 0 \quad \text{as } j \to +\infty, \tag{4.3}$$

where

$$\operatorname{dist}(A, M) = \sup_{a \in A} \inf_{y \in M} \|a - y\|.$$

In order to find such a z, we first study properties of the "limit set" Λ defined by

$$\Lambda := \bigcap_{q \geq 1} cl \bigcup_{n \geq q} O(p_n, F_n). \tag{4.4}$$

LEMMA 4.1. *Under the above assumptions and notation the following properties hold:*

(i) Λ *is a nonempty compact set and* $\Lambda \subset B$;

(ii) Λ *is invariant under* F_0;

(iii) $\operatorname{dist}(O(p_n, F_n), \Lambda) \to 0$ *as* $n \to +\infty$.

Proof. As mentioned in §2, (H)(a) implies that $F : J \times X \to X$ is compact. Therefore, $clF([-\delta, \delta] \times B)$ is a compact set in X. Since $O(p_n, F_n) \subset B$ and, by periodicity, $O(p_n, F_n) = F_n(O(p_n, F_n))$, we have

$$\bigcup_{n \geq 0} O(p_n, F_n) \subset F([-\delta, \delta] \times B).$$

This implies that $cl \bigcup_{n \geq q} O(p_n, F_n), q \geq 1$, is a nested family of compact sets and, as such, it has nonempty compact intersection. Since all the orbits $O(p_n, F_n)$ are contained in B and B is closed, Λ is also contained in B. This proves (i).

To prove (ii), first observe that Λ can be equivalently defined by

(4.5)
$$\Lambda = \{z \in X : \text{there are two sequences } n_j, m_j \text{ of positive integers such that } n_j \to \infty \text{ and } F_{n_j}^{m_j}(p_{n_j}) \to z\}.$$

The verification of (4.5) is left to the reader. (The reader has certainly noticed the analogy to the proof of properties of ω-limit sets of orbits.)

We prove that $F_0(\Lambda) \subset \Lambda$ (positive invariance). Let $z \in \Lambda$, and let n_j, m_j be the sequences as in (4.5). We prove that $F_0(z) \in \Lambda$ by showing that

$$F_{n_j}^{m_j+1}(p_{n_j}) \to F_0(z) \quad \text{as } j \to +\infty.$$

This follows from the triangle inequality and the following convergence properties:

$$F_0 F_{n_j}^{m_j}(p_{n_j}) - F_0(z) \to 0$$

and

$$F_{n_j}^{m_j+1}(p_{n_j}) - F_0 F_{n_j}^{m_j}(p_{n_j}) = F_{n_j} F_{n_j}^{m_j}(p_{n_j}) - F_0 F_{n_j}^{m_j}(p_{n_j}) \to 0.$$

The former convergence is just by continuity of F_0, while the latter one follows from the fact that all points $F_{n_j}^{m_j}(p_{n_j})$ lie in the compact set $clF([-\delta,\delta]\times B)$. Note that it is a straightforward consequence of the continuity of $F : [-\delta,\delta]\times X \to X$ that

$$F_n(z) = F(\varepsilon_n, z) \to F_0(z) \quad \text{as } n \to \infty$$

and that the convergence is uniform for z in a compact set.

To prove that $\Lambda \subset F_0(\Lambda)$ (negative invariance) one proceeds as follows. Given $z \in \Lambda, n_j, m_j$ being the corresponding sequences, one considers the sequence

$$F_{n_j}^{m_j-1}(p_{n_j}).$$

It is well defined since p_{n_j} is periodic. This sequence lies in a compact set, hence passing to a subsequence one may assume that it converges to a $y \in \Lambda$. It is then easy to see that $F_0(y) = z$. This completes the proof of (ii).

Property (iii) follows directly from (4.5) and the fact that $O(p_n, F_n), n = 1, 2, \ldots,$ are contained in a compact set. □

Our next aim is to find a linearly stable periodic point of F_0 in Λ.

First we find a $\tilde{z} \in \Lambda$ with $\lambda_1(\tilde{z}, F_0) \leq 0$. Existence of such a $\tilde{z}$ follows easily by property (iii) of Lemma 4.1 (in case of nonexistence, Lemma 3.1 applied to $K = \Lambda$ would give a contradiction to (4.2)).

Next, consider the orbit $O(\tilde{z}, F_0)$. By invariance of Λ, we have $O(\tilde{z}, F_0) \subset \Lambda$. So $O(\tilde{z}, F_0)$ is relatively compact, and consequently the ω-limit set $\omega(\tilde{z}, F_0)$ is nonempty, compact and invariant under F_0. Moreover,

$$\text{dist}(F_0^n(\tilde{z}), \omega(\tilde{z}, F_0)) \to 0 \quad \text{as } n \to +\infty \tag{4.6}$$

(see, e.g., [Ha]). Obviously,

$$\omega(\tilde{z}, F_0) \subset clO(\tilde{z}, F_0) \subset \Lambda.$$

Now, since

$$\lambda_1(F_0^n(\tilde{z}), F_0) = \lambda_1(\tilde{z}, F_0) \leq 0, \tag{4.7}$$

there is a $z \in \omega(\tilde{z}, F_0)$ with $\lambda_1(z, F_0) \leq 0$ (otherwise Lemma 3.1 applied to $K = \omega(\tilde{z}, F_0)$ gives a contradiction to (4.7)).

In fact, there must be a *regular* point $z \in \omega(\tilde{z}, F_0)$ with $\lambda_1(z, F_0) \leq 0$ (see [PT1, Prop. 4.6]). Proposition 4.4 of [PT1] now yields that $\omega(\tilde{z}, F_0) = O(z, F_0)$ and z is a periodic point of F_0.

We have thus shown that Λ contains a linearly stable periodic point of F_0. To complete the localization we need the following result.

LEMMA 4.2. *Let z be a linearly stable periodic point of F_0. Then for any neighborhood V of $O(z, F_0)$ there exist constants $\rho > 0$ and $\delta_0 > 0$ such that for any $\varepsilon \in [-\delta_0, \delta_0]$ and for any $y \in X$ satisfying $\|y - z\| < \rho$ one of the properties holds:*

(i) *$O(y, F_\varepsilon) \subset V$; or*

(ii) *There are positive integers r, m such that $F_\varepsilon^{r+m} y \gg F_\varepsilon^r y$.*

Before giving the proof we show how this lemma implies localization. Let $z \in \Lambda$ be a linearly stable periodic point of F_0. Then $O(z, F_0) \subset \Lambda$, by invariance of Λ. Further, by (4.5), there are sequences $n_j \to +\infty$ and m_j such that

$$F_{n_j}^{m_j}(p_{n_j}) \to z.$$

Let V be any neighbourhood of $O(z, F_0)$, and let ρ, δ_0 be as in Lemma 4.2. For j sufficiently large (so that $\|F_{n_j}^{m_j}(p_{n_j}) - z\| < \rho$ and $\varepsilon_{n_j} \in [-\delta_0, \delta_0]$) we obtain by Lemma 4.2 that either

$$O(p_{n_j}, F_{n_j}) = O(F_{n_j}^{m_j}(p_{n_j}), F_{n_j}) \subset V \tag{4.8}$$

or else

$$F_{n_j}^{r+m} p_{n_j} \gg F_{n_j}^{r} p_{n_j} \tag{4.9}$$

for some $r, m > 0$. But we know that (4.9) is impossible because the cycle $O(p_{n_j}, F_{n_j})$ cannot contain two related points (cf. §3). Hence (4.8) holds for all j sufficiently large. Since the neighborhood V is arbitrary, this readily completes localization. It remains to prove Lemma 4.2.

Proof of Lemma 4.2. First we show that it is sufficient to consider the case when z is a fixed point of F_0. Indeed, replacing F_ε by F_ε^k, k being the period of z, we still have (H) satisfied and, obviously, z is a linearly stable fixed point of F_0^k. Now, if V is a given neighborhood of $O(z, F_0)$, there exist a $\delta_2 > 0$ and a neighborhood $\tilde{V}$ of z such that $w \in \tilde{V}$ and $\varepsilon \in [-\delta_2, \delta_2]$ imply that $F_\varepsilon^j(w) \in V$ for $j = 0, 1, \dots, k-1$. From this it is clear that if we find $\rho > 0$ and $\tilde{\delta}_0$ corresponding to $\tilde{V}$ as required in Lemma 4.2 with F_ε replaced by F_ε^k, then the same ρ and $\delta_0 = \min\{\delta_2, \tilde{\delta}_0\}$ are the constants required in the original assertion. We thus proceed in the proof assuming that $F_0 z = z$.

Let a neighborhood V of z be given. We want to find constants ρ and δ_0. In order to see how small these constants must be, we first derive an equation for the deviation from z of an orbit $O(y, F_\varepsilon)$ with some fixed y and ε. Denote

$$\begin{aligned} y_0 &= y, \\ y_n &= F_\varepsilon^n y, \\ u_n &= y_n - z. \end{aligned}$$

We have

$$\begin{aligned} u_{n+1} &= F_\varepsilon y_n - F_0 z \\ &= F_0 y_n - F_0 z + F_\varepsilon y_n - F_0 y_n; \end{aligned}$$

hence

$$u_{n+1} = A u_n + g(u_n) + H(\varepsilon, u_n), \tag{4.10}$$

where

$$\begin{aligned} A &= F_0'(z), \\ g(u) &= \int_0^1 [F_0'(z + su) - F_0'(z)] u \, ds, \end{aligned} \tag{4.11}$$

$$H(\varepsilon, u) = F_\varepsilon(u+z) - F_0(u+z). \tag{4.12}$$

The idea of what follows is to show that if u_0 is close to zero and ε is small, then either u_n stays small, so that y_n stays in V, or else the "positive component" of u_n gets so large compared with u_0 that u_n and u_0 are related (hence y_n and y_0 are related). This requires some estimates in the flavor of [PT1, Proof of Prop. 3.5]. First we need some properties of spectral projections corresponding to the Krein–Rutman decomposition of the spectrum of A. By the latter we mean the decomposition

$$\sigma(A) = \{\mu_1\} \cup \sigma_2,$$

where $\mu_1 = \exp(\lambda_1(z, F_0))$ is the principal eigenvalue of (the strongly positive and compact operator) A and $\sigma_2 = \sigma(A)\backslash\mu_1$. The corresponding invariant decomposition of X has the form

$$X = X_1 \oplus X_2, \tag{4.13}$$

where X_1 is spanned by an eigenvector $v \gg 0$, which we further assume to be normalized, $\|v\| = 1$, and $X_2 \cap X_+ = \{0\}$. Moreover, the spectrum of the restriction $A_2 := A|_{X_2}$ is contained inside the unit circle in the complex plane. We can thus find constants $M \geq 1, \beta \in (0,1)$ such that

$$\|A_2^n\| \leq M\beta^n, \qquad n = 0, 1, 2, \ldots. \tag{4.14}$$

Let $P : X \to X_1$ and $Q = I - P : X \to X_2$ be the spectral projections associated with the decomposition (4.13). For $u \in X$ we write

$$u^1 = Pu, \qquad u^2 = Qu.$$

The following property holds true. Let $\mathbb{B}_R$ denote the ball of radius R around zero in X.

SUBLEMMA. *There exist constants $C_1 > 0$ and $\alpha \in (0,1)$ such that for any $u \in X$ and any $R > 0$ the relations*

$$\|u^2\| \leq C_1\|u^1\| \quad \textit{and} \quad \|u\| \geq R \tag{4.15}$$

imply that u is strongly related to $\mathbb{B}_{\alpha R}$, that is, either $u \gg \mathbb{B}_{\alpha R}$ or $u \ll \mathbb{B}_{\alpha R}$.

Here $u \gg \mathbb{B}_{\alpha R}$ means that $u \gg b$ for any $b \in \mathbb{B}_{\alpha R}$ (and similarly for the reversed sign). The proof of the sublemma will be given later.

Equipped with this sublemma, we can now find constants ρ and δ_0 corresponding to the given neighbourhood V of z such that the conclusion of Lemma 4.2 holds.

First we choose a number $R \in (0, 1]$ so small that the following properties hold.

(a) For any $u \in X$ with $\|u\| < R$ one has $z + u \in V$.

(b) There is a constant C_2 such that

$$\left\|\frac{\partial F(\varepsilon, z+u)}{\partial \varepsilon}\right\| \leq C_2 \quad \text{for any } u \text{ with } \|u\| \leq R \quad \text{and} \quad \varepsilon \in [-\delta, \delta]. \tag{4.16}$$

(To achieve this, one just uses the fact that F is C^1 and $[-\delta, \delta]$ is compact.)

(c) Certain estimates (see (4.25) below), which involve only C_1 from the sublemma and certain constants depending only on the operator A and the function g (see 4.11) are satisfied by R. Having such an R we choose a ρ with

$$0 < \rho < \min\{\alpha R, [\|Q\| 3M(1 + C_1^{-1})]^{-1} R\}. \tag{4.17}$$

To define the constant δ_0, we first observe that, by (4.12),

$$h(\varepsilon) := \sup_{\|u\|\le R} \|H(\varepsilon,u)\| \le \varepsilon \sup_{\|u\|\le R} \left\| \frac{\partial F(\varepsilon, z+u)}{\partial \varepsilon} \right\|; \tag{4.18}$$

hence, by (4.16),

$$h(\varepsilon) \to 0 \quad \text{as } \varepsilon \to 0 .$$

Choose $\delta_0 > 0$ so small that $\varepsilon \in [-\delta_0, \delta_0]$ implies

$$h(\varepsilon) M \|Q\| \sum_{k=0}^{\infty} \beta^k < \overline{R}/3. \tag{4.19}$$

Here $M \ge 1$ and $\beta \in (0,1)$ are as in (4.14) and

$$\overline{R} := (1 + C_1^{-1})^{-1} R.$$

With ρ and δ_0 chosen in this way, we claim that if $\varepsilon \in [-\delta_0, \delta_0]$ and $\|u_0\| = \|y - z\| < \rho$, then the following alternative holds (with the notation as above). Either

(i)′ $\|u_n\| < R, n = 0, 1, 2, \ldots$, or else

(ii)′ there is an n such that $\|u_n\| \ge R$ and $\|u_n^2\| \le C_1 \|u_n^1\|$.

By the choice of R, (i)′ implies assertion (i) of Lemma 4.2. Since $\|u_0\| = \|y - z\| < \rho < \alpha R$, (ii)′ implies (ii) (here we use the sublemma). So we are done, provided we prove that the alternative holds. To do that we assume that (ii)′ does not hold, i.e., for any n,

$$\|u_n^1\| < C_1^{-1} \|u_n^2\| \quad \text{or} \quad \|u_n\| < R. \tag{4.20}$$

We prove that (i)′ holds. By (4.20) and the inequality $\|u_n\| \le \|u_n^1\| + \|u_n^2\|$ it is sufficient to prove that

$$\|u_k^2\| < \overline{R} = (1 + C_1^{-1})^{-1} R \tag{4.21}$$

for $k = 0, 1, 2, \ldots$. Observe that (4.21) holds true for $k = 0$ by the choice ρ. Indeed,

$$\|u_0^2\| \le \|Q\| \, \|u_0\| < \|Q\| \rho < (3M)^{-1} \overline{R} < \overline{R}. \tag{4.22}$$

Suppose (4.21) holds for $k = 0, 1, \ldots, n-1$. We prove that it holds for $k = n$. This implies that (4.22) holds for all $n = 0, 1, 2, \ldots$.

Applying the projection Q to both sides of (4.10) we obtain

$$u_{n+1}^2 = A_2 u_n^2 + Q[g(u_n) + H(\varepsilon, u_n)].$$

The variation of constants for this equation gives

$$u_n^2 = A_2^n u_0^2 + \sum_{k=0}^{n-1} A_2^{n-k-1} Q[g(u_k) + H(\varepsilon, u_k)]. \tag{4.23}$$

By (H), the function $\|g(u)\| \, \|u\|^{-1-\theta}$ is bounded on any bounded set. So there is a constant C_3 such that

$$\sup_{\|u\|\le R} \|g(u)\| \le C_3 R^{1+\theta}. \tag{4.24}$$

(Since R is in the bounded interval $[0,1]$, C_3 depends only on g.)

Now we have prepared everything for an estimate of $\|u_n^2\|$. By (4.23) and (4.14)

$$\|u_n^2\| \le M\beta^n\|u_0^2\| + \sum_{k=0}^{n-1} M\beta^{n-k-1}\|Q\|\,(\|g(u_k)\| + \|H(\varepsilon, u_k)\|).$$

By our induction hypotheses (cf. (4.20)) and (4.21) we have $\|u_k\| < R$ for $k = 0, 1, \dots n-1$. We can thus apply (4.18), (4.19), and (4.24), which yield

$$\|u_n^2\| \le M\|u_0^2\| + R^{1+\theta}M\|Q\|C_3\sum_{k=0}^{\infty}\beta^k + \frac{\overline{R}}{3}.$$

Recall that, by the choice of ρ (see (4.22))

$$M\|u_0^2\| < \frac{\overline{R}}{3}.$$

So if R is so small that

$$(1 + C_1^{-1})R^{\theta}M\|Q\|C_3\sum_{k=0}^{\infty}\beta^k < \frac{1}{3}, \tag{4.25}$$

we have the estimate

$$\|u_n^2\| < \overline{R}.$$

This completes the proof of the alternative. It remains to prove the sublemma.

Proof of the sublemma. Define C_1 to be a constant such that

$$v + x \gg \frac{1}{2}v \quad \text{for any } x \quad \text{with } \|x\| \le C_1. \tag{4.26}$$

Such C_1 exists because the eigenvector $\frac{1}{2}v$ is in $\operatorname{int}X_+$. Put $\alpha = C_1(1+C_1)^{-1}$. Observe that if $\|u\| \ge R$ and $\|u^2\| \le C_1\|u^1\|$, then

$$R \le \|u^1 + u^2\| \le (1 + C_1)\|u^1\|;$$

hence

$$\|u^1\| \ge R(1 + C_1)^{-1}.$$

Suppose that (4.15) holds. For definiteness also assume that $u^1 > 0$ (the case $u^1 < 0$ is analogous). Then by (4.15) and (4.26),

$$\begin{aligned} u = u^1 + u^2 &= \|u^1\|v + u^2 = \|u^1\|(v + \|u^1\|^{-1}u^2) \\ &\ge R(1 + C_1)^{-1}\tfrac{1}{2}v. \end{aligned}$$

Now, by (4.26),

$$\tfrac{1}{2}v \gg \mathbb{B}_{C_1}.$$

Therefore,

$$u \ge R(1 + C_1)^{-1}\tfrac{1}{2}v \gg R(1 + C_1)^{-1}\mathbb{B}_{C_1} = \mathbb{B}_{\alpha R}. \qquad \square$$

Having completed localization we pass to the second step.

II. Local bifurcation. Our aim here is to prove the following proposition.

PROPOSITION 4.3. *Let z be a linearly stable k-periodic point of F_0. Then there exist a $\delta_2 > 0$ and a neighborhood V of $O(z, F_0)$ such that for any $\varepsilon \in (-\delta_2, \delta_2)$ all cycles of F_ε contained in V have the minimal period k.*

Proof. By continuity of F, there are $\tilde{\delta}_2 > 0$ and a neighborhood $\tilde{V}$ of $O(z, F_0)$ such that for any $\varepsilon \in (-\tilde{\delta}_2, \tilde{\delta}_2)$ all cycles of F_ε contained in $\tilde{V}$ have the minimal period at least k. Next we prove that such cycles have the minimal period at most k, if $\tilde{\delta}_2$ and $\tilde{V}$ are made sufficiently small.

Similarly as in Lemma 4.2, we reduce the proof to the case when z is a fixed point of F_0. To see that it is sufficient to consider this case, one only has to realize that the following consequence of continuity of F holds: Let $\tilde{V}$ be a given neighborhood of the k-periodic point z such that $F_0^j z \notin cl\tilde{V}$ for $j = 1, 2, \dots, k-1$. Then there exist a $\delta_3 > 0$ and a neighborhood V of $O(z, F_0)$, such that for any $\varepsilon \in (-\delta_3, \delta_3)$ and any cycle $O(y, F_\varepsilon)$ of F_ε contained in V, the set $O(y, F_\varepsilon) \cap \tilde{V}$ is a cycle for F_ε^k. With this observation it is easy to see that the assertion of the lemma follows provided we prove the assertion with F_ε replaced by F_ε^k, assuming that z is a fixed point of F_0^k. Notice that F_ε^k also satisfies (H). Thus we may proceed, assuming that z is a fixed point of F_0.

Consider the spectrum $\sigma(A)$ of the (compact strongly positive) operator $A := F_0'(z)$. We have

$$\sigma(A) = \{\mu_1\} \cup \sigma_2,$$

where $\mu_1 = \exp(\lambda_1(z, F_0))$ and $\sigma_2 \subset \{\mu \in \mathbb{C} : |\mu| < \mu_1\}$. We distinguish between the two cases

$$\mu_1 < 1 \quad \text{and} \quad \mu_1 = 1.$$

(Recall that $\mu_1 \leq 1$, since z is linearly stable.) If $\mu_1 < 1$, there is an equivalent norm $\|\cdot\|_a$ on X such that $\|A\|_a < 1$ (see, e.g., [Io, Chap. I]). Since F is C^1, there exists a ball $\mathbb{B}$ around z in $(X, \|\cdot\|_a)$ such that for all ε sufficiently close to zero, F_ε is a contraction on $\mathbb{B}$. Therefore, all cycles of F_ε in $\mathbb{B}$ are fixed points of F_ε. In the case $\mu_1 < 1$ we are done. Let us now consider the case $\mu_1 = 1$. To prove the assertion, we apply a center manifold theorem. This will give us a neighborhood V of z such that for each ε close to zero, any invariant set of F_ε contained in V lies on a one-dimensional locally invariant manifold of F_ε. Moreover, the restriction of F_ε to this invariant manifold is conjugate to an increasing map on an interval. There are thus no periodic points of F_ε, except for fixed points, on this invariant manifold and consequently there are no periodic points of higher period in the neighborhood V. This implies the assertion.

We give more details concerning the reduction result (which certainly is familiar to experts in the bifurcation theory).

To find the invariant manifolds for all ε, we consider the standard suspension, i.e., the mapping

$$G : (\varepsilon, x) \mapsto (\varepsilon, F(\varepsilon, x)) : (-\delta, \delta) \times X \to (-\delta, \delta) \times X$$

(cf. [He, §8.5]; [Io, Chap. V]).

By (H), G is of class C^1. Obviously, $(0, z)$ is a fixed point of G. The spectrum of the linearized (at $0, z)$) operator $(\eta, w) \mapsto (\eta, d_z F_0 w + \eta d_0 F(\cdot, z))$ is the same as the spectrum of $A = d_z F_0$, except that 1 is now a double eigenvalue (it has geometric multiplicity 2 if $d_0 F(\cdot, z) \in R(I - A)$, the range of $I - A$, otherwise it has geometric multiplicity 1 and algebraic multiplicity 2). The eigenspace corresponding to the eigenvalue 1 has the form $\operatorname{span}\{(1, w), (0, v)\}$, where v is the positive eigenvector of A

and w is a vector in X. By the center manifold theorem (see [He], [Io], [Sh], [MM], [CL]) there exist a $\delta_2 > 0$, a neighborhood V of z, and a C^1 submanifold of $\mathbb{R} \times X$ of the form

$$W := \{(\varepsilon, z + \varepsilon w + sv) + (\sigma_1(\varepsilon, s), \sigma_2(\varepsilon, s)) : \varepsilon, s \in (-\delta_2, \delta_2)\},$$

where σ_1 and σ_2 take values in $\mathbb{R}$ and X, respectively, are of class C^1 and of order $O(|\varepsilon| + |s|)$ as $\varepsilon \to 0, s \to 0$, and the following properties hold: For any $(\varepsilon, x) \in W$ the iterates $G^n(\varepsilon, x), n = 0, 1, \ldots$ stay in W as long as they stay in $(-\delta_2, \delta_2) \times V$. Moreover, any cycle of G contained entirely in $(-\delta_2, \delta_2) \times V$ is contained in W.

Let us be more precise about the references. The center manifold theorem in the form we need is stated in [Io, Chap. V] and [He, §8.5]. In [He] the application of this theorem to parametrized mappings (using the suspension as above) is also described (cf. also [Io, §V.3], where $F(\varepsilon, z) \equiv F(z)$ for any ε is assumed). However, in this references higher regularity of F is required. One has to combine these results with [Sh, Thm. 3.2], where C^1-regularity of the (global) center-unstable manifold (in our case just the center manifold) for a C^1-mapping is provided (see also [CL], which is slightly less general).

Now the special form of G implies that for $\varepsilon \in (-\delta_2, \delta_2)$ the ε-section of W, $W_\varepsilon = \{x : (\varepsilon, x) \in W\} = \{z + \varepsilon w + sv + \sigma_2(\varepsilon, s) : |s| < \delta_2\}$, is a one-dimensional C^1-submanifold of X, which is locally invariant under F_ε (more precisely, $F_\varepsilon(x) \in W_\varepsilon$ if $x \in W_\varepsilon$ and $F_\varepsilon(x) \in V$) and which contains any cycle of F_ε lying in V. Now, since the critical eigenvalue which gives rise to these invariant manifolds (that is, the eigenvalue of $d_z F_0$ on the unit circle) equals 1, by continuous dependence of the eigenvalues on parameters, we may assume (making δ_2 and V smaller if necessary) that the restriction $F_{\varepsilon|W_\varepsilon}$ is orientation preserving. As such, this restriction cannot have periodic orbits, except fixed points. We conclude that for $\varepsilon \in (\delta_2, \delta_2)$, any cycle of F_ε contained in V is a fixed point of F_ε. This completes the proof of Proposition 4.3. □

III. Completion of the proofs. First we complete the proof of Theorem 2. Let B be a closed bounded set. Let m be a bound on stable periods of $F_{0|B}$. Suppose the conclusion of Theorem 2 is not true. Then there exist sequences $\varepsilon_n \in [-\delta, \delta], p_n \in X, k_n \in \{1, 2, \ldots\}$ such that $\varepsilon_n \to 0, p_n$ is a linearly stable k_n-periodic point of $F_n = F_{\varepsilon_n}$ with $O(p_n, F_n) \subset B$ and $k_n > m, n = 1, 2, \ldots$. Using localization and Proposition 4.3 (where z is the k-periodic point found by localization, hence $O(z, F_0) \subset B$ and $k \leq m$ by assumption) we immediately arrive at a contradiction.

Next, suppose that for a given closed bounded set B, the conclusion of Theorem 1 fails. Since there is no bound on stable periods at $F_{0|B}$, there is a sequence p_n of linearly stable periodic points of F_0 with $O(p_n, F_0) \subset B$ such that the minimal period k_n of p_n approaches $+\infty$ as $n \to +\infty$. Using localization (with $\varepsilon_n \equiv 0$) and Proposition 4.3 we again arrive at a contradiction.

5. Applications. There are several standard classes of periodically forced equations whose period maps have the smoothness, compactness, and strong monotonicity properties we require of F_0 (cf. [Hi2], [Hs2]). Moreover, in many cases the backward uniqueness guarantees that the period map is one-to-one. Thus Theorem 1 and Corollary 4 are applicable for a general scalar semilinear or quasilinear equation or a cooperative system of reaction diffusion equations. We do not consider these general classes here. Our concern in this section is to show some applications of Corollaries 3 and 5. We consider three types of problems where nonexistence of linearly stable *subharmonic* solutions (i.e., solutions whose minimal period is a nontrivial multiple of the period of the equation) can be established. All these equations are small perturbations of problems for which the nonexistence is known to hold.

Example 1 (perturbation of an autonomous equation). Consider the parabolic equation

$$u_t = \Delta u + f(x,u) + \varepsilon g(\varepsilon, t, x, u), \qquad t > 0, \quad x \in \Omega, \tag{5.1}$$

where $\Omega \subset \mathbb{R}^N$ is a smooth bounded domain, $\varepsilon \in \mathbb{R}$ a parameter, and the nonlinearities $f : \overline{\Omega} \times \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \times \mathbb{R} \times \overline{\Omega} \times \mathbb{R} \to \mathbb{R}$ satisfy the following assumptions.

(A1) f is of class C^2.

(A2) g is of class C^1.

(A3) g is τ-periodic in t, $\tau > 0$, i.e. $g(\cdot, t+\tau, \cdot, \cdot) \equiv g(\cdot, t, \cdot, \cdot)$.

(A4) There are constants $C_1, C_2 > 0$ such that

$$|g(\varepsilon, t, x, u)| \leq C_1|u| + C_2$$

for any ε, t, x, u.

(A5) $L := \limsup_{|n| \to \infty} f(x,u)/u < \mu_1$, uniformly for $x \in \overline{\Omega}$. Here μ_1 is the first eigenvalue of $-\Delta$ on Ω under the boundary condition

$$\mathcal{B}u = 0, \tag{5.2}$$

where the boundary operator $\mathcal{B}$ is either of Dirichlet type

$$\mathcal{B}u = u|\partial\Omega$$

or of Neumann type

$$\mathcal{B}u = \frac{\partial u}{\partial n} + bu$$

($b : \partial\Omega \to \mathbb{R}^+$ being a C^2-function and n a smooth outward pointing and nowhere tangent vector field on $\partial\Omega$).

We consider the boundary value problem (5.1), (5.2) in the context of analytic semigroups.

Let $Y = L_p(\Omega)$ with $p > N$, and let $Y^\beta, \beta \geq 0$, be the scale of fractional power spaces associated with the L_p-realization of $-\Delta$ and the boundary condition (5.2) (cf. [He]). Choose a β with $(p+N)(2p)^{-1} < \beta < 1$. Then $X := Y^\beta$ is continuously imbedded in $C^1(\overline{\Omega})$; hence it is strongly ordered by the pointwise ordering (see [Hi3], [Hs2]). By [He], (5.1), (5.2) is well posed on X. More specifically, (A1), (A2) imply that for any $u_0 \in X$, (5.1), (5.2) has a unique maximal solution $t \mapsto u(t, \cdot, \varepsilon, u_0)$ in $C([0,s), X) \cap C^1([0,s), X)$ satisfying $u(0, \cdot, \varepsilon, u_0) = u_0(\cdot)$. By (A4), (A5) this solution is global, that is, $s = +\infty$ [Am1], so the period map $F_\varepsilon : u_0 \mapsto F_\varepsilon(u_0) := u(\tau, \cdot, \varepsilon, u_0) : X \to X$ is defined everywhere on X. It is well known (see [He], [Hs2]) that all the smoothness, compactness, and monotonicity properties required in (H) are satisfied by the family $F_\varepsilon; \varepsilon \in [-\delta, \delta]$ (for any $\delta > 0$). Finally, by backward uniqueness, F_ε is one-to-one. Thus (H)(b) is satisfied for F_0 (and also for F_ε if g is assumed more regular, say C^2).

Now fix a $\delta > 0$ such that $\delta C_1 + L < \mu_1$. We want to apply Corollary 3 to the family $F_\varepsilon, \varepsilon \in [-\delta, \delta]$. For any $\varepsilon \in [-\delta, \delta]$ we have

$$\limsup_{|u| \to \infty} (f(x,u) + \varepsilon g(\varepsilon, t, x, u))/u \leq |\varepsilon| C_1 + L \leq \delta C_1 + L < \mu_1,$$

uniformly for $x \in \overline{\Omega}, t \in [0, \tau]$.

This estimate implies that there is a constant $C_3 > 0$ such that for each $\varepsilon \in [-\delta, \delta]$ all periodic solutions of (5.1), (5.2) satisfy the a priori estimate

$$\|u(t, \cdot, \varepsilon, u_0)\|_X \leq C_3.$$

Indeed, using standard comparison arguments (cf. [KW], [Am2]) one constructs a time-independent subsolution in $-\text{int}X_+$ and, similary, a supersolution in $\text{int}X_+$. This can be done in such a way that any multiple by a scalar factor ≥ 1 is again a sub/ supersolution. We thus have unbounded continua of subsolutions and supersolutions. The standard sweeping principle implies a uniform a priori bound on periodic solutions in L_∞. By bootstrap arguments [Am2], we obtain the desired a priori estimate in X.

Thus for $\varepsilon \in [-\delta, \delta]$ all the cycles of F_ε lie in a ball D independent of ε.

Finally, since F_0 is the time-τ map of a C^1-strongly monotone semiflow, all linearly stable periodic points of F_0 are fixed points of F_0 corresponding to equilibria of the autonomous problem (see [Hs2, Prop. 9.4]).

We have thus verified all the hypotheses of Corollary 3. We infer that for $|\varepsilon|$ sufficiently small there are no linearly stable periodic points of F_ε except for fixed points. Consequently, (5.1), (5.2) has no stable subharmonic solutions. If in addition to (A1)–(A5) one assumes that $g(\varepsilon, t, x, u)$ is of class C^2, then the F_ε also satisfy the hypotheses of Corollary 5. As a consequence we obtain that for $|\varepsilon|$ sufficiently small, the following convergence property holds: For any bounded set $\mathcal{G} \subset X$ there is an open dense subset $\tilde{\mathcal{G}}$ of $\mathcal{G}$ such that $O(u_0, F_\varepsilon)$ is convergent for any $u_0 \in \tilde{\mathcal{G}}$.

Various generalizations of this example are possible. Firstly, the Laplacian may be replaced by a general uniformly elliptic operator of second order with C^2 coefficients. Assuming appropriate growth conditions one may allow f and g to depend also on ∇u.

We now mention two further examples. Without going into details, we present periodically time-dependent equations which do not admit linearly stable subharmonic orbits, and discuss admissible perturbations of such equations.

Example 2 (perturbation of a spatially homogeneous equation). Let Ω be a smooth convex domain in $\mathbb{R}^N$. Consider the problem

$$\begin{aligned} &u_t = \Delta u + f(t, u, \nabla u), \qquad t > 0, \quad x \in \Omega, \\ &\frac{\partial u}{\partial n} = 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{5.3}$$

where n is now the unit outward normal vectorfield on $\partial\Omega$. The function $f : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ is assumed to be of class C^2 and τ-periodic in t. It has been observed in [PT2] that there are no linearly stable subharmonic solutions. This follows from the fact that any stable $k\tau$-periodic solution, $k \in \{1, 2, \ldots\}$, is spatially homogeneous (see [Hs2, Thm. 23.4]). Therefore, it satisfies the τ-periodic one-dimensional ODE

$$u_t = f(t, u, 0),$$

which does not have subharmonic solutions.

Consider now the perturbed equation

$$u_t = \Delta u + f(t, u, \nabla u) + \varepsilon g(t, x, u, \nabla u) \tag{5.4}$$

with a small spatial inhomogeneity (ε is a small parameter). Under appropriate smoothness and growth conditions one can again check that the period map F_ε on X

(with X as in Example 1) satisfies (H). Thus there are no stable subharmonic solutions of (5.3), (5.4) for small $|\varepsilon|$, and "most" solutions converge to a τ-periodic solution.

Example 3 (perturbation in one space dimension). Consider the one-dimensional problem

$$u_t = u_{xx} + f(t, x, u, u_x), \qquad t > 0, \quad 0 < x < 1, \tag{5.5}$$

$$u_x(t, 0) \equiv u_x(t, 1) \equiv 0, \tag{5.6}$$

where f is of class C^2 and τ-periodic in t (other boundary conditions are also admissible). It has been proved in [BPS] that any bounded solution approaches, as $t \to \infty$, a τ-periodic solution. In particular, there is no subharmonic solution.

We now perturb (5.5) by a nonlocal term,

$$u_t = u_{xx} + f(t, x, u, u_x) + \varepsilon^2 C(x) \int_0^1 \nu(x) u(t, x) dx, \tag{5.7}$$

where $C(x) \geq 0$ and $\nu(x) \geq 0$ are C^1 functions of $x \in [0, 1]$. These sign restrictions assure that the strong comparison principle applies to (5.6) and (5.7). Under appropriate growth conditions the period map of this equation satisfies (H) with $X = H^2(0, 1) \cap H_0^1(0, 1)$ (see [PT2]). Using Corollary 3, one obtains that there are no stable subharmonic solutions of (5.6), (5.7) if ε is sufficiently small. By [PT1] this implies that "most" orbits of the period map are convergent.

REFERENCES

[AH1] N. D. Alikakos and P. Hess, *On stabilization of discrete monotone dynamical systems*, Israel J. Math., 59 (1987), pp. 185–194.

[AH2] ———, *Liapunov operators and stabilization in strongly order preserving dynamical systems*, Differential Integral Equations, 4 (1991), pp. 15–24.

[AHM] N. D. Alikakos, P. Hess, and H. Matano, *Discrete order preserving semigroups and stability for periodic parabolic differential equations*, J. Differential Equations, 82 (1989), pp. 322–341.

[Am1] H. Amann, *Global existence for semilinear parabolic systems*, J. Reine Angew. Math., 366 (1985), pp. 47–89.

[Am2] ———, *Periodic solutions of semilinear parabolic equations*, in Nonlinear Analysis, L. Cesari, R. Kannan, and H. F. Weinberger, eds., Academic Press, New York, 1978.

[BPS] P. Brunovský, P. Poláčik, and B. Sandstede *Convergence in general periodic parabolic equations in one space dimension*, Nonlinear Anal., 18 (1992), pp. 209–215.

[CL] S.-N. Chow and K. Lu, *C^k centre unstable manifolds*, Proc. Royal Soc. Edinburgh, 108A (1988), pp. 303–320.

[DH] E. N. Dancer and P. Hess, *Stability of fixed points for order preserving discrete-time dynamical systems*, J. Reine Angew. Math., 419 (1991), pp. 125–139.

[Ha] J. K. Hale, *Functional Differential Equations*, Springer-Verlag, New York, 1977.

[He] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, New York, 1981.

[Hs1] P. Hess, *On stabilization of discrete strongly order-preserving semigroups and dynamical processes*, in Semigroup Theory and Applications, Ph. Clément et al., eds., Dekker L.N. 116, 1989, pp. 231–240.

[Hs2] ———, *Periodic-parabolic boundary value problems and positivity*, Pitman Research Notes in Mathematics 247, Longman Scientific and Technical, New York, 1991.

[Hi1] M. W. Hirsch, *Differential equations and convergence almost everywhere in strongly monotone flows*, Contemp. Math., 17 (1983), pp. 267–285.

[Hi2] M. W. HIRSCH, *Attractors for discrete-time dynamical systems in strongly ordered spaces*, in Geometry and Topology, J. Alexander, J. Harer, eds., Springer-Verlag, Berlin, Heidelberg, 1985.

[Hi3] ———, *Stability and convergence in strongly monotone dynamical systems*, J. Reine Angew. Math., 383 (1988), pp. 1–58.

[Hi4] ———, *Systems of differential equations that are competitive or cooperative* IV: *Structural stability in three-dimensional systems*, SIAM J. Math. Anal., 21 (1991), pp. 1225–1234.

[Io] G. IOOSS, *Bifurcation of Maps and Applications*, North-Holland, Amsterdam, 1979.

[KW] J. L. KAZDAN AND F. W. WARNER, *Remarks on some quasilinear elliptic equations*, Comm. Pure Appl. Math., 28 (1975), pp. 567–597.

[M1] R. MAÑE, *Ergodic Theory and Differentiable Dynamics*, Spinger-Verlag, Berlin, Heidelberg, 1987.

[M2] ———, *Liapunov exponents and stable manifolds for compact transformations*, in Geometric Dynamics, J. Palis, ed., Springer-Verlag, New York, 1983, pp. 522–577.

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

[Ma] H. MATANO, *Strong comparison principle in nonlinear parabolic equations*, in Nonlinear Parabolic Equations: Qualitative Properties of Solutions, L. Boccardo and A. Tesei, eds., Pitman, London, 1987, pp. 148–155.

[Mi1] J. MIERCZYŃSKI, *On generic behavior in strongly cooperative differential equations*, Colloq. Math. Soc. János. Bolyai, 53 (1990), pp. 402–406.

[Mi2] ———, *Flows on ordered bundles*, preprint.

[P1] P. POLÁČIK, *Convergence in smooth strongly monotone flows defined by semilinear parabolic equations*, J. Equations, 79 (1989), pp. 89–110.

[P2] ———, *Generic properties of strongly monotone semiflows defined by ordinary and parabolic differential equations*, Colloq. Math. Soc. János. Bolyai, 53 (1990), pp. 402–406.

[PT1] P. POLÁČIK AND I. TEREŠČÁK, *Convergence to cycles as a typical asymptotic behavior in smooth strongly monotone discrete-time dynamical systems*, Arch. Rational Mech. Anal., 116 (1991), pp. 339–360.

[PT2] ———, *Exponential separation and invariant bundles for maps in ordered Banach spaces, with applications in parabolic equations*, J. Dynamics Differential Equations, to appear.

[Sh] M. SHUB, *Global Stability of Dynamical Systems*, Springer-Verlag, New York, 1987.

[Sm] 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.

[ST1] H. L. SMITH AND H. R. THIEME, *Quasiconvergence and stability for strongly order-preserving semiflows*, SIAM J. Math. Anal., 21 (1990), pp. 673–692.

[ST2] ———, *Convergence for strongly order-preserving semiflows*, SIAM J. Math. Anal., 22 (1991), pp. 1081–1101.

[T1] P. TAKÁČ, *Convergence to equilibrium on invariant d-hypersurfaces for strongly increasing discrete-time dynamical systems*, J. Math. Anal. Appl., 148 (1990), pp. 223–244.

[T2] ———, *Domains of attraction of generic ω-limit sets for strongly monotone semiflows*, Z. Anal. Anwendungen, 10 (1991), pp. 275–317.

[T3] ———, *Asymptotic behavior of strongly monotone time-periodic dynamical processes with symmetry*, J. Diff. Equations, 100 (1992), pp. 355–378.

[T4] ———, *Linearly stable subharmonic orbits in strongly monotone time-periodic dynamical systems*, Proc. Amer. Math. Soc., 115 (1992), pp. 691–698.

[T5] ———, *Domains of attraction of generic ω-limit sets for strongly monotone discrete-time semigroups*, J. Reine Angew. Math., 423 (1992), pp. 101–173.

SIAM J. MATH. ANAL.
Vol. 24, No. 5, pp. 1331–1339, September 1993

012

STRICTLY NONAUTONOMOUS COOPERATIVE SYSTEM WITH A FIRST INTEGRAL*

BAORONG TANG†§, YANG KUANG†‡, AND HAL SMITH†§

Abstract. The authors consider the nonautonomous cooperative system

$$dx_i/dt = F_i(t, x_1, \dots, x_n) \qquad (i = 1, \dots, n)$$

in the nonnegative orthant in the real n-dimensional Euclidean space, which has the first integral with positive gradient. The authors guess that every solution to such a system either converges to an asymptotic state (for the almost periodic (or periodic) case, this state is an almost periodic (or periodic) solution) or eventually leaves any compact set. They partly prove this conjecture.

Key words. cooperative systems, nonautonomous systems, almost periodic (or periodic) solution, first integral, Lyapunov function, uniformly stable, skew-product flow, ω-limit set

AMS subject classifications. primary 34C27; secondary 34C25, 34D20, 90A16

1. Introduction.

The autonomous gross-substitute system forms a mathematical model for the classical law of supply and demand in economics, which has been studied by many economists [9] and has the property (P_1): *every bounded solution converges to an equilibrium.* Such a system is a special class of cooperative systems.

Cooperative (or competitive) systems are a class of dynamic systems $\dot{x} = F(x)$, $x \in U \subset \mathbf{R}^n$, which satisfy conditions $\partial F_i/\partial x_j \geq 0$ (≤ 0) for $i \neq j$. Recently, systems of this type received much study (see references in [3]–[5] and [15]), since they have wide applications in biology and chemistry (for example, see [6], [17]). In [3]–[5] Hirsch proved many important results about such systems, one of which is that, for systems that are cooperative and irreducible, almost all (with respect to the Lebesgue measure) points whose forward orbits are bounded approach the equilibrium set. But in general, the property (P_1) that every bounded solution converges to an equilibrium or to a closed orbit is not true. A construction of Smale [13] shows that any dynamics is possible for such systems. Hence, one tries to find some special classes of systems of that type for which the property (P_1) is true. In [14] Smillie established a particular class of cooperative systems for which the property (P_1) is true. In [7] Mierczyński found another class of cooperative systems, namely, strictly cooperative systems with a first integral, having the property (P_1) which generalized the result in [9]. Other relevant results can be found in Arino [19] and the references cited therein.

As we know, nonautonomous systems are more realistic and more general in mathematical modelling. For example, if one considers the seasonal effects in economics, it is important to study time-dependent or nonautonomous gross-substitute systems, which are a special class of nonautonomous cooperative systems. For the case of periodic (almost periodic) cooperative systems, the property (P_2) that *every bounded*

*Received by the editors September 3, 1991; accepted for publication (in revised form) December 3, 1992.

†Department of Mathematics, Arizona State University, Tempe, Arizona 85287-1804.

‡The research of this author was partially supported by National Science Foundation grant DMS 9102549.

§The research of this author was partially supported by National Science Foundation grant DMS 89-22654.

solution converges to a periodic (almost periodic) solution is important. Smith [16] generalized the result in [14] to the periodic case: the property (P_2) holds. Nakajima [8] and Sell and Nakajima [12] studied the nonautonomous gross-substitute systems for which the property (P_2) is true.

In this paper we study the periodic (almost periodic) cooperative systems with a general class of first integrals. For such systems, whether the property (P_2) is true remains unknown. We conjecture here that such a property continued to hold.

We partially answer the above conjecture in this work. Under some conditions, we are able to show that this conjecture is indeed true. Our results generalize those of [8] and [12]. Without loss of generality, we concentrate our study on the almost periodic case.

We adapt the idea similar to that in [12]: by using a Lyapunov function, we show that any "positively compact" solution of the above system (defined in §2) is asymptotically almost periodic. The first integral used in [8] and [12] is $\sum_{i=1}^{n} x_i$, for which there are many properties which played a very important role in [8] and [12]. But since we consider more general first integrals, these properties fail to be true. As we shall see, we introduce a new class of first integrals to construct proper Lyapunov functions and apply the theory of cooperative systems to obtain the desired results. These Lyapunov functions play important roles in this work.

The paper is organized as follows. In the next section, we present some definitions and preliminary lemmas. In §3 we prove the main result, which partly gives an affirmative answer to the conjecture. In the last section we give an example.

2. Definitions and preliminary lemmas. Let $\mathbf{R}^n$ denote the real n-dimensional Euclidean space with norm $|x| = \sum_{i=1}^{n} |x_i|$ for $x = (x_1, \dots, x_n) \in \mathbf{R}^n$ and set $\mathbf{R}_+ = \{x \in \mathbf{R} : x \geq 0\}$, $\mathbf{R}_+^n = \{x \in \mathbf{R}^n : x_i \geq 0\}$, $\partial\mathbf{R}_+^n = \{x \in \mathbf{R}_+^n : x_i = 0 \text{ for some } i\}$ and $\operatorname{Int}\mathbf{R}_+^n = \mathbf{R}_+^n \backslash \partial\mathbf{R}_+^n$.

We denote $x < y$ if $x_i < y_i$ for each i, and $x \leq_i y$ if $x \neq y$, $x_i = y_i$ and $x_j \leq y_j$ for $j \neq i$. By $\langle \cdot, \cdot \rangle$, we mean the usual inner product in $\mathbf{R}^n$.

Let $H : \mathbf{R}_+^n \to \mathbf{R}$ be a C^1 function. We denote grad $H(p)$ as the *gradient* of H at p, which is the vector $((\partial H/\partial x_1)(p), \dots, (\partial H/\partial x_n)(p))$. We denote Int $H^{-1}(h)$ as the set $\{x \in \operatorname{Int}\mathbf{R}_+^n : H(x) = h\}$.

Let $F : \mathbf{R} \times \mathbf{R}_+^n \to \mathbf{R}^n$ be a vector field. A *first integral* for F is a function $H : \mathbf{R}_+^n \to \mathbf{R}_+$, of class C^1, such that grad $H(p) \neq 0$ at each $p \in \mathbf{R}_+^n$ and $\langle \operatorname{grad} H(p), F(t,p) \rangle \equiv 0$.

The system of ordinary differential equations we consider takes the form

$$\frac{dx_i}{dt} = F_i(t, x_1, \dots, x_n) = F_i(t,x), \qquad x \in \mathbf{R}_+^n,\ 1 \leq i \leq n, \tag{2.1}$$

where $F(t,x) = (F_1(t,x), \dots, F_n(t,x))$ is defined and C^1 on $\mathbf{R} \times \mathbf{R}_+^n$. Throughout this paper, we assume that $F(t,x)$ satisfies the following conditions

(A1) If $x \leq_i y$ then $F_i(t,x) < F_i(t,y)$;

(A2) There exists a first integral H for F such that $\operatorname{grad} H(x) > 0$ for $x \in \mathbf{R}_+^n$ and $H(0) = 0$;

(A3) $F(t,x)$ is uniformly almost periodic in t; or

(A3)′ $F(t,x)$ is periodic in t with period $\omega > 0$.

Remark. (A1) implies that (2.1) is cooperative. Also, the assumption that $H(0) = 0$ is not necessary.

Define the translate F_τ by $F_\tau(x,t) = F(x, \tau + t)$, where $\tau \in \mathbf{R}$. By the *hull*, we mean the set $\mathcal{F} = \operatorname{Cl}\{F_\tau : \tau \in \mathbf{R}\}$, where the closure is taken in the topology of uniform convergence on compact sets. It is known that $\mathcal{F}$ is an almost periodic minimal set [2], [11], [18]. It is easy to see that every $G \in \mathcal{F}$ satisfies the conditions (A2), (A3) and that if $x \leq_i y$ then $G(t,x) \leq G(t,y)$.

Recall that a mapping $\pi : W \times \mathbf{R} \to W$ is a flow if π is continuous, $\pi(w,0) = w$ and $\pi(\pi(w,s),t) = \pi(w,s+t)$ for all $w \in W$ and $s,t \in \mathbf{R}$, where W is a topological Hausdorff space. If W is a product space $W = X \times Y$, then a flow π is said to be a *skew-product flow* if π has the form $\pi = (\varphi, \sigma)$, or

$$\pi(x,y,t) = (\varphi(x,y,t), \sigma(y,t)),$$

where $\sigma : Y \times T \to Y$ is itself a flow on Y. For each $x \in \mathbf{R}^n_+$ and $G \in \mathcal{F}$, we denote by $\varphi(x,G,t)$ the maximally defined solution of $x' = G(x,t)$ that satisfies $\varphi(x,G,0) = x$. It is known that

$$\pi(x,G,\tau) = (\varphi(x,G,\tau), G_\tau) \tag{2.2}$$

describes a (local) skew-product flow on $\mathbf{R}^n_+ \times \mathcal{F}$ [10].

A solution $\varphi(x,F,t)$ of (2.1) is said to be *uniformly stable* if it is defined for all $t \geq 0$ and for every $\epsilon > 0$ there is a $\delta = \delta(\epsilon) > 0$ such that

$$|\varphi(x,F,\tau+t) - \varphi(y,F,\tau+t)| \leq \epsilon \quad \text{for all} \quad t \geq 0$$

whenever $\tau \geq 0$ and $|\varphi(x,F,\tau) - \varphi(y,F,\tau)| \leq \delta$ [10].

Let $V(x,y) : \mathbf{R}^n_+ \times \mathbf{R}^n_+ \to \mathbf{R}$ satisfy the following conditions.

(i) $V(x,y) > 0$ for all $x,y \in \mathbf{R}^n$ with $x \neq y$;

(ii) $V(x,y) = 0$ if and only if $x = y$;

(iii) $\lim_{|x-y| \to +\infty} V(x,y) = +\infty$.

Then we say that V is *positively definite.*

We shall use the following lemma, which is easily verified.

LEMMA 2.1. *Let $\varphi(\hat{x},F,t)$ be a positively compact solution, i.e., $\varphi(\hat{x},F,t)$ remains in a compact subset of $\mathbf{R}^n$ for all $t \geq 0$. Assume that there is a positively definite function $V(x,y)$ on $\mathbf{R}^n_+ \times \mathbf{R}^n_+$ such that for all $x,y \in \mathbf{R}^n_+$ and $G \in \mathcal{F}$, one has $D^+V(\varphi(x,G,t),\varphi(y,G,t)) \leq 0$, where D^+ denotes the right-hand derivative. Then $\varphi(\hat{x},F,t)$ is uniformly stable.*

Theorems 2 and 5 in [10] yields the following theorem.

THEOREM 2.1. *Let π be the skew-product flow* (2.2) *on $\mathbf{R}^n_+ \times \mathcal{F}$ generated by system* (2.1). *Let $\varphi(\hat{x},F,t)$ be a positively compact uniformly stable solution of* (2.1) *and let Ω denote the ω-limit set of the motion $\pi(\hat{x},F,t)$. Then Ω is a nonempty compact connected distal minimal set. Furthermore, if for some $G \in \mathcal{F}$ the section*

$$\Omega(G) = \{x \in \mathbf{R}^n_+ : (x,G) \in \Omega\}$$

has only finitely many points, then Ω is an almost periodic minimal set, and for each $(x,G) \in \Omega$ the solution $\varphi(x,G,t)$ is almost periodic in t.

The definition of a distal set can be found in [10].

Recall that a compact invariant set M is minimal if and only if every trajectory is dense in M. If Ω is minimal, then for $x,y \in \Omega(F)$, $(x,F) \in \Omega$, $(y,F) \in \Omega$. There is a sequence $t_n \to +\infty$ such that $\varphi(x,F,t_n) \to y$. $F_{t_n} \to F$. Since $\{\varphi(y,F,t_n)\}$ is compact, without loss of generality, suppose that $\varphi(y,F,t_n) \to z$. Now $F_{t_n} \to F$ and

$(z, F) \in G$. Hence, if Ω is minimal, then for $x, y \in \Omega(F)$, there is a sequence $t_n \to +\infty$ such that $\varphi(x, F, t_n) \to y$ and $\varphi(y, F, t_n) \to z$, where $z \in \Omega(F)$.

Let $x(t) = \varphi(x, G, t)$, $y(t) = \varphi(y, G, t)$ be two solutions of the system $x' = G(t, x)$. Suppose that they are defined on a common interval I. For $t \in I$, we define the following two subsets of $\{i : 1 \leq i \leq n\}$:

$$P_t = \{i : x_i(t) \geq y_i(t)\}, \qquad Q_t = \{i : x_i(t) \leq y_i(t)\}.$$

3. Partial answer.

In this section we study system (2.1) which has $H(x)$ as a first integral.

In the rest of this paper, we make the following assumption:

$$\text{(A4)} \qquad \frac{\partial H}{\partial x_i \partial x_j} \geq 0 \quad \text{for} \quad i \neq j.$$

Define the function $V(x, y) : \mathbf{R}^n_+ \times \mathbf{R}^n_+ \to \mathbf{R}_+$ as follows:

$$V(x, y) = \sum_{i=1}^{n} |H(x_1, \dots, x_{i-1}, x_i, y_{i+1}, \dots, y_n) - H(x_1, \dots, x_{i-1}, y_i, y_{i+1}, \dots, y_n)|.$$

Let

$$V^+(x, y) = \sum_{i \in P_t} [H(x_1, \dots, x_{i-1}, x_i, y_{i+1}, \dots, y_n) - H(x_1, \dots, x_{i-1}, y_i, y_{i+1}, \dots, y_n)],$$

$$V^-(x, y) = \sum_{i \in Q_t} [H(x_1, \dots, x_{i-1}, y_i, y_{i+1}, \dots, y_n) - H(x_1, \dots, x_{i-1}, x_i, y_{i+1}, \dots, y_n)],$$

then $V(x, y) = V^+(x, y) + V^-(x, y)$.

LEMMA 3.1. *Let $x(t) = \varphi(x, F, t)$, $y(t) = \varphi(y, F, t)$ be two solutions of the system* (2.1) *with a common interval I, then for $t_1, t_2 \in I$ with $t_1 < t_2$, we have*

$$V(x(t_2), y(t_2)) \leq V(x(t_1), y(t_1)).$$

Proof. We first prove that

$$\text{(3.1)} \qquad V^+(x(t_2), y(t_2)) \leq V^+(x(t_1), y(t_1)).$$

Let $z_i(t_1) = \max\{x_i(t_1), y_i(t_1)\}$, $z(t_1) = (z_1(t_1), \dots, z_n(t_1))$, then $z(t_1) \geq y(t_1)$ and $z(t_1) \geq x(t_1)$. Let $z(t) = \varphi(z(t_1), F, t - t_1)$ be a solution of (2.1), then (A1) and Kamke's theorem (e.g., see [2]–[4] and [15]) yield that

$$z(t) \geq y(t) \quad \text{and} \quad z(t) \geq x(t) \quad \text{for } t \geq t_1,$$

i.e., $z(t) \geq \max\{x(t), y(t)\} = (\max\{x_1(t), y_1(t)\}, \dots, \max\{x_n(t), y_n(t)\})$ for $t \geq t_1$. Thus, $V^+(z(t), y(t)) = V(z(t), y(t)) = H(z_1(t), \dots, z_n(t)) - H(y_1(t), \dots, y_n(t))$ for $t \geq t_1$. Since H is the first integral for F, we have

$$\text{(3.2)} \qquad V(z(t), y(t)) = V(z(t_1), y(t_1)) = V^+(x(t_1), y(t_1)) \text{ for } t \geq t_1.$$

Since $z(t_2) \geq \max\{x(t_2), y(t_2)\}$, the assumption (A4) implies that

$$\text{(3.3)} \qquad V(z(t_2), y(t_2)) \geq V(\max\{x(t_2), y(t_2)\}, y(t_2)) = V^+(x(t_2), y(t_2)).$$

By (3.2) and (3.3), (3.1) follows.

Similar arguments show that

$$V^-(x(t_2), y(t_2)) \leq V^-(x(t_1), y(t_1))$$

and the proof is thus completed. □

Remark. Clearly, if $V(x, y)$ defined above satisfies that $\lim_{|x-y| \to +\infty} V(x, y) = +\infty$, then V is positively definite.

LEMMA 3.2. *In* (2.1), *in addition to* (A1), (A2), (A3) (*or* (A3)′), *and* (A4), *assume that* $\lim_{|x-y|\to+\infty} V(x,y) = +\infty$ *and for every* $h \in \mathbf{R}_+$, *there is at most one bounded solution which is defined and belongs to* $H^{-1}(h)$ *for all* $t \in \mathbf{R}$. *Then for every positively compact solution* $\hat{x}(t)$, *there exists an almost periodic solution* (*or periodic solution*) *such that*

$$\lim_{t\to+\infty} |\hat{x}(t) - \phi(t)| = 0. \tag{3.4}$$

Proof. Let $\hat{x}(t) = \varphi(\hat{x}, F, t)$ be a positively compact solution and Ω be the ω-limit set of the corresponding motion $\pi(\hat{x}, F, t)$ in $\mathbf{R}^n_+ \times \mathcal{F}$. Let $\Omega(F) = \{x \in \mathbf{R}^n_+ : (x, F) \in \Omega\}$, then Lemma 3.1, combining Lemma 2.1 and Theorem 2.1, implies that $\Omega(F)$ consists of a single point y and the solution $\phi(t) = \varphi(y, F, t)$ is almost periodic.

Now we prove (3.4).

By Lemma 3.1, one has $D^+V(\hat{x}(t), \phi(t)) \leq 0$ for all $t \geq 0$. Thus, the limit $\lim_{t\to+\infty} V(\hat{x}(t), \phi(t))$ exists. Choose a sequence $t_n \to +\infty$ so that $\pi(\varphi(y, F, t_n)) \to (y, F)$ and $\varphi(\hat{x}, F, t_n) \to z$. Since $F_{t_n} \to F$ it follows that $z \in \Omega(F)$ and so $z = y$. Thus $\lim_{t\to+\infty} V(\hat{x}(t), \phi(t)) = 0$ and so $\lim_{t\to+\infty} |\hat{x}(t) - \phi(t)| = 0$. □

For $y = (y_1, \dots, y_n) \in \mathbf{R}^n$, we define

$$S_k(y) = \{x = (x_1, \dots, x_n) \in \mathbf{R}^n : x_i \leq y_i,\ i = 1, \dots, k-1,\ x_k < y_k \text{ and } x_{k+1} > y_{k+1},\ x_i \geq y_i \text{ for } i = k+2, \dots, n\},$$

$$k = 1, \dots, n-1.$$

LEMMA 3.3. *Let* $x(t), y(t)$ *be two solutions of* (2.1). *If* $x(t) \in \cup_{k=1}^{n-1} S_k(y(t))$, *then*

$$\frac{dV(x(t), y(t))}{dt} < 0.$$

Proof. We consider the case $x(t) \in S_1(y(t))$; for the other cases, the proofs are similar.

If $x(t_0) \in S_1(y(t_0))$, then there is a $\delta > 0$ such that $x(t) \in S_1(y(t))$ for $t_0 - \delta < t < t_0 + \delta$. Hence, for $t \in (t_0 - \delta, t_0 + \delta)$,

$$\begin{aligned} V(x(t), y(t)) = {} & H(y_1(t), \dots, y_n(t)) + H(x_1(t), \dots, x_n(t)) \\ & - 2H(x_1(t), y_2(t), \dots, y_n(t)). \end{aligned}$$

Since $H(x_1(t), \dots, x_n(t)) = H(y_1(t), \dots, y_n(t)) = \text{constant}$, we have

$$\begin{aligned} \frac{dV(x(t), y(t))}{dt} = {} & -2\frac{\partial H(x_1(t), y_2(t), \dots, y_n(t))}{\partial x_1} F_1(t, x_1(t), \dots, x_n(t)) \\ & - 2\sum_{i=2}^{n} \frac{\partial H(x_1(t), y_2(t), \dots, y_n(t))}{\partial y_i} F_i(t, y_1(t), \dots, y_n(t)). \end{aligned}$$

Notice that

$$\begin{aligned} & \frac{\partial H(x_1(t), y_2(t), \dots, y_n(t))}{\partial x_1} F_1(t, x_1(t), y_2(t), \dots, y_n(t)) \\ & + \sum_{i=2}^{n} \frac{\partial H(x_1(t), y_2(t), \dots, y_n(t))}{\partial y_i} F_i(t, x_1(t), y_2(t), \dots, y_n(t)) = 0 \end{aligned}$$

and so

$$\begin{aligned}\frac{dV(x(t),y(t))}{dt} = &-2\frac{\partial H(x_1(t),y_2(t),\dots,y_n(t))}{\partial x_1}\\ &\times (F_1(t,x_1(t),x_2(t),\dots,x_n(t)) - F_1(t,x_1(t),y_2(t),\dots,y_n(t)))\\ &-2\sum_{i=2}^{n}\frac{\partial H(x_1(t),y_2(t),\dots,y_n(t))}{\partial y_i}\\ &\times (F_i(t,y_1(t),y_2(t),\dots,y_n(t)) - F_i(t,x_1(t),y_2(t),\dots,y_n(t))).\end{aligned}$$

Now assumption (A1) yields that

$$\frac{dV(x(t),y(t))}{dt} < 0. \qquad \square$$

For any two points $x, y \in \mathbf{R}^n$, which are not related, i.e., that $x \le y$ or $y \le x$ is not true, we always find a map $P : \mathbf{R}^n \to \mathbf{R}^n$ such that $Px \in S_k(y)$ for some $k \in \{1,\dots,n-1\}$.

Let $\mathcal{T}$ be the set of such maps. Since $\mathbf{R}^n$ is finite dimensional, $\mathcal{T}$ is finite. Define

$$V_P(x,y) = V(Px,Py) \quad \text{for} \quad P \in \mathcal{T},$$

and let $W(x,y) = \sum_{P\in\mathcal{T}} V_P(x,y)$, then we have

LEMMA 3.4. *Let $x(t), y(t)$ be two solutions of* (2.1) *with $H(x(t)) = H(y(t))$. Then*

$$\frac{dW(x(t),y(t))}{dt} < 0.$$

Proof. Let $P \in \mathcal{T}$. Consider the system

$$\frac{du}{dt} = F_P(t,u), \tag{3.5}$$

where $u = Px$, $F_P(t,u) = P(F(t,P^{-1}u))$. It is easy to check that system (3.5) satisfies (A1) to (A4). By Lemma 3.1 and Lemma 3.2, we have

$$D^+V(Px(t),Py(t)) \le 0$$

and if $Px(t) \in \cup_{k=1}^{n-1} S_k(Py(t))$, then $\frac{dV(Px(t),Py(t))}{dt} < 0$. Hence

$$D^+V_p(x(t),y(t)) \le 0$$

and

$$\frac{dV_p(x(t),y(t))}{dt} < 0 \quad \text{for} \quad Px(t) \in \bigcup_{k=1}^{n-1} S_k(Py(t)).$$

For $x(t), y(t)$ with $H(x(t)) = H(y(t))$, (A2) implies that $x(t)$ and $y(t)$ are not related and so we can choose $P_0 \in \mathcal{T}$ such that $P_0x(t) \in S_k(P_0y(t))$ for some k, then the above argument shows that

$$\frac{dV_{P_0}(x(t),y(t))}{dt} < 0.$$

Now

$$\frac{dW(x(t),y(t))}{dt} = \sum_{P\in\mathcal{T}} \frac{dV_P(x(t),y(t))}{dt} < 0. \qquad \square$$

We are now in a position to prove our main result.

THEOREM 3.1. *In* (2.1), *in addition to* (A1), (A2), (A3) (*or* (A3)′), *and* (A4), *assume that* $\lim_{|x-y|\to\infty} V(x,y) = +\infty$. *Then for every positively compact solution* $\hat{x}(t)$, *there exists an almost periodic solution* (*or periodic solution*) *such that*

$$\lim_{t\to+\infty} |\hat{x}(t) - \phi(t)| = 0.$$

Proof. Let $\hat{x}(t) = \varphi(\hat{x}, F, t)$ be a positively compact solution and Ω the ω-limit set of the corresponding motion $\pi(\hat{x}, F, t)$ in $\mathbf{R}^n_+ \times \mathcal{F}$. Let $\Omega(F) = \{x \in \mathbf{R}^n_+ : (x, F) \in \Omega\}$. Suppose $\Omega(F)$ contains more than one point. Let $x_0, y_0 \in \Omega(F)$ with $x_0 \neq y_0$ and $x(t) = \varphi(x_0, F, t)$, $y(t) = \varphi(y_0, F, t)$ be the corresponding solution of (2.1).

Since $x(t)$ and $y(t)$ stay in a compact set in $\mathbf{R}^n_+$, they are defined for all $t \in \mathbf{R}$.

By Theorem 2.1, Ω is a distal set. Ellis's theorem [1] implies that the product flow on $\Omega \times \Omega$ is the union of minimal sets. Hence there is a sequence $t_n \to +\infty$ such that $x(t_n) \to x_0$ and $y(t_n) \to y_0$.

Lemma 3.4 implies that there is a $\tau \in \mathbf{R}$ (say $\tau > 0$) such that

$$W(x(t), y(t)) \leq W(x(\tau), y(\tau)) < W(x_0, y_0) \quad \text{for} \quad t \geq \tau.$$

Letting $t_n \to +\infty$, one gets the contradiction

$$W(x_0, y_0) = \lim_{t_n\to+\infty} W(x(t_n), y(t_n)) < W(x_0, y_0).$$

For $\tau \leq 0$, a suitable translation of $x(t)$ and $y(t)$ also yields a contradiction. Thus, the proof follows from Lemma 3.2. □

Remark. If we take $H(x_1, \ldots, x_n) = x_1 + \cdots + x_n$, it is trivial that (A4) is satisfied. Hence the above result generalizes those of [8] and [12].

In the rest of this section, we suppose that (2.1) is periodic, i.e., (A3)′ is true and (A1), (A2) and (A4) hold. Define the Poincaré map $T : \mathbf{R}^n_+ \to \mathbf{R}^n_+$ as follows

$$Tx = \varphi(x, F, \omega).$$

For $x_0 \in \mathbf{R}^n_+$, define $x_0 + \mathbf{R}^n_+ = \{x_0 + x : x \in \mathbf{R}^n_+\}$. Then it is easy to prove the following lemma.

LEMMA 3.5. *Let* $x_0 \in \operatorname{Int} \mathbf{R}^n_+$ *be a periodic point, i.e., the solution* $\varphi(x_0, F, t)$ *is periodic with the period* ω. *Then* $x_0 + \mathbf{R}^n_+$ *is positively invariant under* T, *i.e., for* $x \in (x_0 + \mathbf{R}^n_+) - x_0$, $Tx \in \operatorname{Int}(x_0 + \mathbf{R}^n_+)$. *Thus,* x_0 *is a unique periodic point on* $x_0 + \partial\mathbf{R}^n_+$.

LEMMA 3.6. *Let* $x_0 \in \operatorname{Int} \mathbf{R}^n_+$ *be a periodic point. Then, for every* $\epsilon > 0$ *there exists* $\delta > 0$ *such that for each* $h \in [H(c) - \delta, H(c) + \delta] \cap [0, M)$, *where* M *is the least upper bound of the values of* H, *there is a periodic point* x_h *such that* $H(x_h) = h$, $x_h > x_0$ *for* $h > H(x_0)$ (*respectively,* $x_h < x_0$ *for* $h < H(x_0)$) *and* $|x_h - x_0| < \epsilon$.

Proof. We consider the case $h > H(x_0)$. Let $e_i = (0, \ldots, 0, {}_i1, 0, \ldots, 0)$, then from (A2), we have $H(x_0 + e_i) > H(x_0)$. Let $\delta_1 = \min_{1\leq i\leq n}\{H(x_0 + e_i) - H(x_0)\}$, then (A2) implies that $H^{-1}(h) \cap (x_0 + \rho e_i) \neq \phi$ for all $1 \leq i \leq n$, $\rho \geq 0$, and $h \in [H(x_0), H(x_0) + (\delta_1/2)]$. It is easy to see that $H^{-1}(h) \cap (x_0 + \mathbf{R}^n_+)$ is homeomorphic to the $(n-1)$-dimensional disk. Lemma 3.2 and the definition of H shows that T maps $H^{-1}(h) \cap (x_0 + \mathbf{R}^n_+)$ to itself. The Brouwer fixed-point theorem yields that there is a periodic point $x_h \in H^{-1}(h) \cap (x_0 + \mathbf{R}^n_+)$ for all $h \in [H(x_0), H(x_0) + (\delta_1/2)]$. Also $x_h > x_0$.

Clearly, we can choose a point $x_1 > x_0$ such that for all $x \in H^{-1}(H(x_0) + (\delta_1/2)) \cap (x_0 + \mathbf{R}^n_+)$, $x < x_1$. Then for all $x \in H^{-1}(h) \cap (x_0 + \mathbf{R}^n_+)$ with $h \in [H(x_0), H(x_0) + (\delta_1/2)]$, we have $x_0 \leq x < x_1$. Let $m = \min\{\partial H(x)/\partial x_i : 1 \leq i \leq n,\ x_0 \leq x \leq x_1\}$. If we choose $\delta < \min\{m\epsilon, (\delta_1/2)\}$, then it is easy to conclude that for $h \in [H(x_0), H(x_0) + \delta) \cap [0, M)$, we have

$$|x - x_0| < \epsilon \text{ for } x \in H^{-1}(h),$$

in particular, $|x_h - x_0| < \epsilon$. □

THEOREM 3.2. *In* (2.1) *assume that the assumptions made in Theorem* 3.1 *are true. Then the set of periodic points in* Int $\mathbf{R}^n_+$ *is connected.*

Proof. It follows from Lemma 3.6 and Theorem 3.1. □

4. An example. Consider the following three-dimensional system in $\mathbf{R}^3_+$:

$$\begin{aligned} \dot{x}_1 &= \frac{a(t)x_2 + b(t)x_3 - 2c(t)x_1}{3 + x_2 + x_3}, \\ \dot{x}_2 &= \frac{c(t)x_1 - 2a(t)x_2 + b(t)x_3}{3 + x_1 + x_3}, \\ \dot{x}_3 &= \frac{c(t)x_1 + a(t)x_2 - 2b(t)x_3}{2 + x_1 + x_2}, \end{aligned} \tag{4.1}$$

where $a(t) > 0$, $b(t) > 0$, $c(t) > 0$, $a(t), b(t)$ and $c(t)$ are almost periodic. Clearly, $H(x_1, x_2, x_3) = (Hx_1)(2 + x_2) + (1 + x_1)(1 + x_3) + (1 + x_2)(2 + x_3) - 5$ is the first integral and system (4.1) satisfies (A1), (A2), (A3), and (A4). Also,

$$V(x, y) = |(x_1 - y_1)(3 + y_2 + y_3)| + |(x_2 - y_2)(3 + x_1 + y_3)| + |(x_3 - y_3)(2 + x_1 + x_2)|$$

is positively definite. Hence, we can apply Theorem 3.1 to system (4.1).

REFERENCES

[1] R. ELLIS, *Distal transformation groups*, Pacific J. Math., 8 (1958), pp. 401–405.

[2] A. M. FINK, *Almost periodic differential equations*, Lecture Notes in Math. 377 (1974), Springer-Verlag, New York.

[3] M. W. HIRSCH, *Systems of differential equations which are competitive or cooperative* I: *Limit sets*, SIAM J. Math. Anal., 13 (1982), pp. 167–179.

[4] ———, *Systems of differential equations that are competitive or cooperative* II: *Convergence almost everywhere*, SIAM J. Math. Anal., 16 (1985), pp. 432–439.

[5] ———, *The dynamical systems approach to differential equations*, Bull. Amer. Math. Soc. (N.S), 11 (1984), pp. 1–64.

[6] D. S. LEVINE, *Qualitative theory of a third order nonlinear system with examples in population dynamics and chemical kinetics*, Math. Biosci., 77 (1985), pp. 17–33.

[7] J. MIERCZYŃSKI, *Strictly cooperative systems with a first integral*, SIAM J. Math. Anal., 18 (1987), pp. 642–646.

[8] F. NAKAJIMA, *Periodic time-dependent gross-substitute systems*, SIAM J. Appl. Math., 36 (1979), pp. 421–427.

[9] F. NIKAIDO, *Convex Structure and Economic Theory*, Academic Press, New York, 1968.

[10] R. J. SACKER AND G. R. SELL, *Lifting properties in skew-product flows with applications to differential equations,*, Mem. Amer. Math. Soc. 190, 1977.

[11] G. R. SELL, *Nonautonomous differential equations and topological dynamics*, Trans. Amer. Math. Soc., 127 (1967), pp. 241–283.

[12] G. R. SELL AND F. NAKAJIMA, *Almost periodic gross-substitute dynamical systems*, Tôhoku Math. J., 32 (1980), pp. 255–263.

[13] S. SMALE, *On the differential equations of species in competition*, J. Math. Biol., 3 (1976), pp. 5–7.

[14] J. SMILLIE, *Competitive and cooperative tridiagonal systems of differential equations*, SIAM J. Math. Anal., 15 (1984), pp. 530–534.

[15] H. L. SMITH, *Systems of ordinary differential equations which generate an order preserving flow. A survey of results*, SIAM Rev., 30 (1988), pp. 87–113.

[16] ———, *Periodic tridiagonal competitive and cooperative systems of differential equations*, SIAM J. Math. Anal., 22 (1991), pp. 1102–1109.

[17] B. TANG, *On the existence of periodic solutions in the limited explodator model for the Belonsov-Zhabotinskii reaction,*, Nonlinear Anal., TMA, 13 (1989), pp. 1359–1374.

[18] T. YOSHIZAWA, *Stability theory and the existence of periodic and almost periodic solutions*, Lectures in Appl. Math., 14, Springer-Verlag, New York, 1975.

[19] O. ARINO, *Monotone semi-flows which have a monotone first integral*, preprint.

SIAM J. MATH. ANAL.
Vol. 24, No. 5, pp. 1340–1354, Sepetember 1993

013

ON THE INSTABILITY OF ARBITRARY BIORTHOGONAL WAVELET PACKETS*

A. COHEN[†] AND I. DAUBECHIES[‡]

Abstract. Starting from a multiresolution analysis and the corresponding orthonormal wavelet basis, Coifman and Meyer have constructed wavelet packets, a library from which many different orthonormal bases can be picked. This paper proves that when the same procedure is applied to biorthogonal wavelet bases, not all the resulting wavelet packets lead to Riesz bases for $L^2(\mathbb{R})$.

Key words. wavelet packets, biorthogonal wavelets

AMS subject classifications. 26A16, 26A18, 26A27, 39B12

1. Short review of orthonormal wavelet bases.

An orthonormal basis of wavelets $\psi_{j,k}(x) = 2^{j/2}\psi(2^jx-k)$, $j,k \in \mathbb{Z}$, associated with a multiresolution analysis, is completely determined by a 2π-periodic function $m_0(\xi)$. More precisely,

$$\hat{\psi}(\xi) = e^{-i\xi/2}\overline{m_0\left(\frac{\xi}{2}+\pi\right)}\hat{\phi}\left(\frac{\xi}{2}\right), \tag{1.1}$$

with

$$\hat{\phi}(\xi) = (2\pi)^{-1/2}\prod_{j=1}^{\infty} m_0(2^{-j}\xi). \tag{1.2}$$

Here $^\wedge$ denotes the Fourier transform, normalized by

$$\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}}\int dx e^{-ix\xi} f(x).$$

Conversely, given a 2π-periodic function m_0, one can define (1.1) and (1.2); if m_0 satisfies a few conditions, then the resulting ψ will generate an orthonormal wavelet basis. Which conditions? Let us assume that m_0 is continuous (which is the case in all useful examples). Then in order for (1.2) to converge, we need $m_0(0) = 1$. If moreover $|m_0(\xi) - 1| \leq C|\xi|^\alpha$ for some $\alpha > 0$, then (1.2) converges uniformly on compact sets. (This is not really necessary, but satisfied in all examples of even the remotest interest.) Furthermore, orthonormality of the $\psi_{j,k}$ implies that

$$|m_0(\xi)|^2 + |m_0(\xi+\pi)|^2 = 1 \tag{1.3}$$

(see Mallat [14]). This is not sufficient to ensure orthonormality of the $\psi_{j,k}$, however; to guarantee this orthonormality we need one more (necessary and sufficient) condition on m_0, of a more technical nature: there should exist a compact set K, congruent with $[-\pi,\pi]$ modulo 2π, such that

$$\inf_{\xi\in K}\inf_{n\geq 1}|m_0(2^{-n}\xi)| > 0. \tag{1.4}$$

* Received by the editors April 27, 1992; accepted for publication November 20, 1992.

[†] Centre de Recherche de Mathématiques de la Décision, Université Paris-Dauphine, France.

[‡] Mathematics Department, Hill Center, Busch Campus, Rutgers University, New Brunswick, New Jersey 08930, and AT&T Bell Laboratories, Murray Hill, New Jersey 07974.

If all these conditions on m_0 are verified, then the $\psi_{j,k}$'s do indeed constitute an orthonormal basis for $L^2(\mathbb{R})$. Note that $m_0(0) = 1$ automatically ensures that $|m_0(2^{-n}\xi)| > \frac{1}{2}$ for sufficiently large n and all $\xi \in K$ (because K is compact), so that (1.4) is a constraint for only finitely many values of n. See Cohen [4] or Cohen [5] for a proof of the necessity and sufficiency of this last condition; the condition can also be recast in other forms [13], [5], [6]. A consequence of (1.4) is that (see Cohen [5])

$$\inf_{\xi \in K} |\hat{\phi}(\xi)| > 0. \tag{1.5}$$

The orthonormal basis $\{\psi_{j,k}; j, k \in \mathbb{Z}\}$ generated by ψ can be interpreted within the framework of a *multiresolution analysis* (see Mallat [14], Meyer [15]). Let V_0 be the space spanned by the functions $\phi(x-k)$, $k \in \mathbb{Z}$ (which are also orthonormal). Define V_j to be the space obtained by dilating V_0 by 2^j,

$$f \in V_j \Longleftrightarrow f(2^{-j}\cdot) \in V_0;$$

an orthonormal basis of V_j is given by $\{\phi_{j,k};\ k \in \mathbb{Z}\}$, with $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$. Then

$$\cdots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots,$$

and

$$\bigcap_{j \in \mathbb{Z}} V_j = \{0\}, \qquad \overline{\bigcup_{j \in \mathbb{Z}} V_j} = L^2(\mathbb{R}).$$

Let W_j be the orthogonal complement in V_{j+1} of V_j. Then $\{\psi_{j,k};\ k \in \mathbb{Z}\}$ is an orthonormal basis in W_j, and

$$\bigoplus_{j=-\infty}^{\infty} W_j = L^2(\mathbb{R}). \tag{1.6}$$

This is the standard decomposition of $L^2(\mathbb{R})$ into different "layers" of wavelets with resolution 2^{-j}. One can also choose to use only reasonably fine scale wavelets, and to lump the coarser aspects together into one space, corresponding to the decomposition

$$L^2(\mathbb{R}) = V_0 \oplus \left(\bigoplus_{j=0}^{\infty} W_j \right). \tag{1.7}$$

2. Orthonormal wavelet packets.

Given a 2π-periodic function m_0 which satisfies all the conditions in §1, one can define many other orthonormal bases, corresponding to decompositions of $L^2(\mathbb{R})$ different from (1.6) or (1.7). They are all designated by the name "wavelet packets," first defined by Coifman and Meyer; for a discussion of their properties and some applications, see the two papers by Coifman, Meyer, and Wickerhauser in [16]. Their construction can be understood easily by using the following lemma [8].

LEMMA 2.1 (the "splitting trick"). *Suppose that the functions $f_k(x) = f(x-k)$, $k \in \mathbb{Z}$, are orthonormal. Define f^0, f^1 by*

$$\hat{f}^\sigma(\xi) = m_\sigma(\xi/2)\hat{f}(\xi/2), \qquad \sigma = 0, 1,$$

where m_0 *is as above,* $m_1(\xi) = e^{-i\xi}\,\overline{m_0(\xi+\pi)}$. *Then the functions* $f_k^0(x) = \frac{1}{\sqrt{2}} f^0\left(\frac{x}{2} - k\right)$, $f_k^1(x) = \frac{1}{\sqrt{2}} f^1\left(\frac{x}{2} - k\right)$, $k \in \mathbb{Z}$, *constitute an orthonormal basis for* $E = \overline{\text{Span}\{f_k\}}$.

Remark. Note that with this notation convention, $f_0^\sigma(x) = \frac{1}{\sqrt{2}} f^\sigma\left(\frac{x}{2}\right)$ and not $f^\sigma(x)$.

Proof.

1. Since

$$\begin{aligned}\langle f_k, f_\ell \rangle &= \int d\xi |\hat{f}(\xi)|^2 e^{i(k-\ell)\xi} \\ &= \int_0^{2\pi} d\xi e^{i(k-\ell)\xi} \sum_{m\in\mathbb{Z}} |\hat{f}(\xi + 2\pi m)|^2,\end{aligned}$$

orthonormality of the f_k is equivalent to $\sum_{m\in\mathbb{Z}} |\hat{f}(\xi + 2\pi m)|^2 = \frac{1}{2\pi}$ almost everywhere.

2. Similarly,

$$\begin{aligned}\langle f_k^0, f_\ell^0 \rangle &= 2\int d\xi |\hat{f}^0(2\xi)|^2 e^{2i(k-\ell)\xi} \\ &= 2\int_0^{\pi} d\xi e^{2i(k-\ell)\xi} \sum_{m\in\mathbb{Z}} |\hat{f}^0(2\xi + 2\pi m)|^2.\end{aligned}$$

Splitting the sum over m into even and odd m leads to

$$\begin{aligned}2\sum_{m\in\mathbb{Z}} |\hat{f}^0(2\xi + 2\pi m)|^2 &= 2|m_0(\xi)|^2 \sum_n |\hat{f}(\xi + 2\pi n)|^2 \\ &\quad + 2|m_0(\xi+\pi)|^2 \sum_n |\hat{f}(\xi + \pi + 2\pi n)|^2 = \frac{1}{\pi},\end{aligned}$$

proving that the f_k^0 are orthonormal. Orthonormality of the f_k^1 is proved analogously, as well as orthogonality of f_k^0 and f_ℓ^1.

3. On the other hand, if $\sum_k c_k f_k \perp f_\ell^\sigma$ for all $\ell \in \mathbb{Z}$ $\sigma = 0, 1$, then

$$0 = \left\langle f_\ell^\sigma, \sum_k c_k f_k \right\rangle = \sqrt{2} \sum_k \bar{c}_k \int_0^{2\pi} d\xi e^{i(k-2\ell)\xi} m_\sigma(\xi),$$

i.e. $c(\xi) = \sum_k c_k e^{-ik\xi}$ is orthogonal (in $L^2([0, 2\pi])$) to the $e^{-i2\ell\xi} m_\sigma(\xi), \ell \in \mathbb{Z}$, $\sigma = 0, 1$, implying

$$\overline{c(\xi)} m_\sigma(\xi) + \overline{c(\xi+\pi)} m_\sigma(\xi+\pi) = 0 \quad \text{a.e.}, j = 0, 1.$$

Multiplying with $\overline{m_\sigma(\xi)}$, and adding the two equations gives

$$\begin{aligned}0 &= \overline{c(\xi)}\,\left[|m_0(\xi)|^2 + m_0(\xi+\pi)|^2\right] \\ &\quad + \overline{c(\xi+\pi)}\,\left[m_0(\xi+\pi)\overline{m_0(\xi)} - \overline{m_0(\xi)} m_0(\xi+\pi)\right] = \overline{c(\xi)},\end{aligned}$$

proving that $\sum_k c_k f_k = 0$, so that the $\{f_\ell^\sigma; \ell \in \mathbb{Z}, \sigma = 0, 1\}$ span all of E. □

Modulo appropriate dilations, the splitting trick can of course also be applied to spaces with an orthonormal basis generated by the regularly spaced translates $f(x - ak)$, $k \in \mathbb{Z}$ of a single function, even if $a \neq 1$.

If the splitting trick is applied to V_0, the space with orthonormal basis $\{\phi_{0,k}; k \in \mathbb{Z}\}$, where $f = \phi$, then the corresponding functions f^0, f^1 are exactly $f^0(x) = \phi(x)$ and $f^1(x) = \psi(x)$; it easily follows that the splitting of V_0 into the spaces generated by the f_k^0 on one hand, the f_k^1 on the other hand, is exactly $V_0 = V_{-1} \oplus W_{-1}$. One can, therefore, view the transition from (1.7) to (1.6) as a result of infinitely many successive splittings, where at every step the V_j space with smallest index gets split into two.

Starting with (1.7) one can of course choose to apply the splitting trick to many subspaces other than V_0, leading to many different orthonormal bases. Every one of the functions generated in this way can be labelled by an overall dilation J, a translation k, and a sequence ϵ consisting of only ones and zeros, and ending in a tail of all zeros. Concretely,

$$f_{J,k;\epsilon}(x) = 2^{J/2}\psi_\epsilon(2^J x - k)$$

with

$$\hat{\psi}_\epsilon(\xi) = (2\pi)^{-1/2}\prod_{j=1}^{\infty} m_{\epsilon_j}(2^{-j}\xi);$$

if $j_{\max}$ is the largest index for which $\epsilon_{j_{\max}} = 1$, this can be rewritten as

$$\hat{\psi}_\epsilon(\xi) = \left[\prod_{j=1}^{j_{\max}} m_{\epsilon_j}(2^{-j}\xi)\right] \hat{\phi}(2^{-j_{\max}}\xi).$$

The function $f_{J,k;\epsilon}$ is the result of $j_{\max} - 1$ splittings of $W_{j_{\max}+J-1}$. For every fixed choice of a sequence of splittings the result is an orthonormal wavelet packet basis. In applications to signal analysis, one can use entropy estimates to find the "best" basis (Coifman and Wickerhauser [11]).

An important special case is where each W_j space is split exactly j times. The resulting orthonormal basis functions are the integer translates of all the ψ_ϵ, with ϵ ranging over all possible sequences of zeros and ones, with a tail of all zeros. This orthonormal basis is, of all the wavelet packet bases, the closest to a windowed Fourier transform.

3. Biorthogonal wavelet bases. There exist orthonormal wavelet bases with compactly supported ψ and ϕ. The 2π-periodic function m_0 is then a trigonometric polynomial. By imposing a factorization of the type

$$m_0(\xi) = \left(\frac{1 + e^{-i\xi}}{2}\right)^N Q(\xi),$$

one can construct ϕ and ψ with arbitrarily high degree of smoothness [12]. One inconvenience of these compactly supported orthonormal wavelets is that they are not symmetric. One can restore symmetry by relaxing the orthonormality requirement. In this case one works with *two* 2π-periodic functions, m_0 and $\tilde{m}_0$, satisfying

$$m_0(\xi)\overline{\tilde{m}_0(\xi)} + m_0(\xi + \pi)\overline{\tilde{m}_0(\xi + \pi)} = 1. \tag{3.1}$$

There are similarly two pairs of scaling functions and wavelets, defined by

$$(3.2)\qquad \hat{\phi}(\xi) = (2\pi)^{-1/2}\prod_{j=1}^{\infty} m_0(2^{-j}\xi), \qquad \hat{\tilde{\phi}}(\xi) = (2\pi)^{-1/2}\prod_{j=1}^{\infty} \tilde{m}_0(2^{-j}\xi)$$

$$\hat{\psi}(\xi) = e^{-i\xi/2}\overline{\tilde{m}_0\left(\tfrac{\xi}{2}+\pi\right)}\hat{\phi}\left(\tfrac{\xi}{2}\right), \qquad \hat{\tilde{\psi}}(\xi) = e^{-i\xi/2}\overline{m_0\left(\tfrac{\xi}{2}+\pi\right)}\hat{\tilde{\phi}}\left(\tfrac{\xi}{2}\right).$$

The $\psi_{j,k}$ and $\tilde{\psi}_{j,k}$ now constitute dual Riesz bases; both ψ and $\tilde{\psi}$ can be symmetric (or antisymmetric) and have compact support (Cohen, Daubechies, and Feauveau [9]). (Note that there also exists a different scheme of biorthogonal Riesz bases of wavelets, in which ψ and $\tilde{\psi}$ are symmetric or antisymmetric, and one of them is compactly supported, while the other is not. See Chui and Wang [3] and Chui [1].)

This corresponds to m_0 and $\tilde{m}_0$ which are trigonometric polynomials. We shall restrict ourselves to this case. In order for the whole construction to work, we need, of course, to impose again some conditions on m_0 and $\tilde{m}_0$. First, we need $m_0(0) = 1 = \tilde{m}_0(0)$; the infinite products in (3.2) then converge uniformly on compact sets. We also need more technical conditions. One of them is similar to the orthonormal case; i.e., we need that for some compact set K, congruent to $[-\pi, \pi]$ modulo 2π,

$$(3.3)\qquad \inf_{\xi\in K}\inf_{n>0}|m_0(2^{-n}\xi)| > 0$$

and

$$\inf_{\xi\in K}\inf_{n>0}|\tilde{m}_0(2^{-n}\xi)| > 0.$$

Another condition concerns the spectral radius of two matrices derived from m_0 and $\tilde{m}_0$ (see Cohen and Daubechies [7]). One consequence of (3.3) (the only one we shall use here) is that

$$\inf_{\xi\in K}|\hat{\phi}(\xi)| > 0, \qquad \inf_{\xi\in K}|\hat{\tilde{\phi}}(\xi)| > 0$$

(see Cohen, Daubechies and Feauveau [9]).

The two pairs of scaling functions generate two multiresolution hierarchies,

$$\cdots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots$$
$$\cdots \subset \tilde{V}_{-2} \subset \tilde{V}_{-1} \subset \tilde{V}_0 \subset \tilde{V}_1 \subset \tilde{V}_2 \subset \cdots,$$

and the $W_j = \overline{\mathrm{Span}\{\psi_{j,k};\ k \in \mathbb{Z}\}}$, (or $\tilde{W}_j = \overline{\mathrm{Span}\{\tilde{\psi}_{j,k};\ k \in \mathbb{Z}\}}$) are still complement spaces of the V_j in V_{j+1} (or $\tilde{V}_j$ in $\tilde{V}_{j+1}$), though no longer orthogonal complements. The two hierarchies are linked via the property that, for all $j \in \mathbb{Z}$,

$$\tilde{W}_j \perp V_j, W_j \perp \tilde{V}_j.$$

We can again decompose $L^2(\mathbb{R})$ as either

$$(3.4)\qquad L^2(\mathbb{R}) = \bigoplus_{j\in\mathbb{Z}} W_j$$

or

$$(3.5)\qquad L^2(\mathbb{R}) = V_0 \oplus \bigoplus_{j=0}^{\infty} W_j,$$

where the direct sums are not sums of orthogonal spaces. If we define Q_j to be the (nonorthogonal) projection operator onto W_j associated with this expansion, then the L^2-norm $\|u\|^2$ is still equivalent with $\sum_j \|Q_j u\|^2$:

LEMMA 3.1. *Let $\psi_{j,k}$, $\tilde{\psi}_{j,k}$ be biorthogonal wavelet bases, as defined above, with*

$$A\sum_{j,k} |c_{j,k}|^2 \le \left\|\sum_{j,k\in\mathbb{Z}} c_{j,k}\psi_{j,k}\right\|^2 \le B\sum_{j,k} |c_{j,k}|^2.$$

Then for all $u \in L^2(\mathbb{R})$,

$$A/B\sum_{j\in\mathbb{Z}} \|Q_j u\|^2 \le \|u\|^2 \le B/A\sum_{j\in\mathbb{Z}} \|Q_j u\|^2.$$

Proof. We can write $u = \sum_{j,k} c_{j,k}\psi_{j,k}$. It then follows that $Q_j u = \sum_k c_{j,k}\psi_{j,k}$, and

$$\begin{aligned}\|u\|^2 &= \left\|\sum_{j,k\in\mathbb{Z}} c_{j,k}\psi_{j,k}\right\|^2 \le B\sum_{j,k} |c_{j,k}|^2 \\ &= B\sum_j \left(\sum_k |c_{j,k}|^2\right) \le B/A\sum_j \left\|\sum_k c_{j,k}\psi_{j,k}\right\|^2 \le B/A\sum_j \|Q_j\psi\|^2.\end{aligned}$$

The lower bound is proved analogously. □

A similar theorem holds, of course, for the splitting (3.5).

4. The biorthogonal splitting trick. A natural question is now whether wavelet packets can be generalized to the biorthogonal setting. Let us first generalize the "splitting trick."

LEMMA 4.1 (the "biorthogonal splitting trick"). *Suppose that the functions $f_k(x) = f(x-k)$ constitute a Riesz basis for their closed linear span E, with*

$$A|c_k|^2 \le \left\|\sum_k c_k f_k\right\|^2 \le B\sum_k |c_k|^2, \tag{4.1}$$

for all square integrable sequences $(c_k)_{k\in\mathbb{Z}}$. Define f^0, f^1 by

$$\hat{f}^\sigma(\xi) = m_\sigma(\xi/2)\hat{f}(\xi/2), \qquad \sigma = 0, 1,$$

with m_0 as above, and $m_1(\xi) = e^{-i\xi}\overline{\tilde{m}_0(\xi+\pi)}$. Then the functions $f_k^0 = \frac{1}{\sqrt{2}}f^0\left(\frac{x}{2}-k\right)$, $f_k^1(x) = \frac{1}{\sqrt{2}}\, f^1\left(\frac{x}{2}-k\right)$, $k \in \mathbb{Z}$ constitute a Riesz basis for E, with

$$A'\sum_k \left[|a_k|^2 + |b_k|^2\right] \le \left\|\sum_k \left[a_k f_k^0 + b_k f_k^1\right]\right\|^2 \le B'\sum_k \left[|a_k|^2 + |b_k|^2\right], \tag{4.2}$$

where

$$\begin{aligned}B' &= B(\max(M_0, \tilde{M}_0) + \Delta_0) + \frac{B-A}{2}(M_0\tilde{M}_0)^{1/2} \\ A' &= \left[A^{-1}(\max(M_0, \tilde{M}_0) + \Delta_0) + \frac{B-A}{2AB}(M_0\tilde{M}_0)^{1/2}\right]^{-1}\end{aligned}$$

and

$$\begin{aligned}
M_0 &= \sup_\xi \left[|m_0(\xi)|^2 + |m_0(\xi+\pi)|^2\right] \\
\tilde{M}_0 &= \sup_\xi \left[|\tilde{m}_0(\xi)|^2 + |\tilde{m}_0(\xi+\pi)|^2\right] \\
\Delta_0 &= \sup_\xi |m_0(\xi)\tilde{m}_0(\xi+\pi) - m_0(\xi+\pi)\tilde{m}_0(\xi)|\,.
\end{aligned}$$

Remark. If the f_k are orthonormal to start with, and if $\tilde{m}_0 = m_0$ (orthonormal filter case), then $M_0 = \tilde{M}_0 = 1$, $\Delta_0 = 0$, $B = A = 1$, and the new bounds B', A' are also equal to 1. The estimates below can also be used to prove bounds of the type (4.2) where the constants B'', A'' are simply proportional to B and A respectively, namely

$$\begin{aligned}
B'' &= B\left[\max(M_0,\tilde{M}_0) + (M_0\tilde{M}_0)^{1/2}\right] \\
A'' &= A\left[\max(M_0,\tilde{M}_0) + (M_0\tilde{M}_0)^{1/2}\right]^{-1}\,;
\end{aligned} \tag{4.3}$$

these "simpler" bounds are less sharp, in the sense that they do not collapse to 1 if everything is orthonormal.

Proof.

1. For $(c_k)_{k\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$, we denote by $c(\xi)$ the 2π-periodic function $c(\xi) = \sum_k c_k e^{-ik\xi} \in L^2([0,2\pi])$. Then

$$\begin{aligned}
\left\|\sum_k c_k f_k\right\|^2 &= \int_{-\infty}^{\infty} d\xi \left|\sum_k c_k e^{-ik\xi}\hat{f}(\xi)\right|^2 \\
&= \int_0^{2\pi} d\xi |c(\xi)|^2 \sum_k |\hat{f}(\xi+2\pi k)|^2,
\end{aligned}$$

so that (4.1) is seen to be equivalent to $A/2\pi \le \sum_k |\hat{f}(\xi+2\pi k)|^2 \le B/2\pi$.

2. Define $\tilde{f} \in E$ by $\langle \tilde{f}, f_k\rangle = \delta_{0,k}$. Then the $\tilde{f}_\ell(x) = \tilde{f}(x-\ell)$, $\ell \in \mathbb{Z}$ constitute the dual Riesz basis for the f_k. In particular,

$$B^{-1}\sum_k |c_k|^2 \le \left\|\sum_k c_k\tilde{f}_k\right\|^2 \le A^{-1}\sum_k |c_k|^2 \tag{4.4}$$

and

$$\langle \tilde{f}_\ell, f_k\rangle = \delta_{k,\ell}.$$

(Both can easily be derived from $\hat{\tilde{f}}(\xi) = \frac{1}{2\pi}\hat{f}(\xi)[\sum_k |\hat{f}(\xi+2\pi k)|^2]^{-1}$.)

3. We start by proving that the f_k^0, f_ℓ^1 span all of E. Assume that $u \in E$ is orthogonal to all the f_ℓ^σ, $\ell \in \mathbb{Z}$, $\sigma = 0, 1$. Since the $\tilde{f}_k$ constitute a Riesz basis for E, we can write $u = \sum_k c_k\tilde{f}_k$. We have then

$$\begin{aligned}
0 = \langle u, f_\ell^\sigma\rangle &= \sqrt{2}\int_{-\infty}^{\infty} d\xi\overline{c(\xi)\hat{\tilde{f}}(\xi)}\hat{f}(\xi)m_\sigma(\xi)e^{-2i\ell\xi} \\
&= \sqrt{2}\int_0^\pi d\xi e^{-2i\ell\xi}\sum_{n\in\mathbb{Z}} m_\sigma(\xi+n\pi)\overline{c(\xi+n\pi)\hat{f}(\xi+n\pi)\hat{\tilde{f}}(\xi+n\pi)}.
\end{aligned}$$

Because m_σ and c are 2π-periodic, and $\sum_\ell \hat{f}(\xi + 2\ell\pi)\overline{\widehat{\tilde{f}}\,(\xi + 2\ell\pi)} = \frac{1}{2\pi}$, this implies

$$m_\sigma(\xi)\overline{c(\xi)} + m_\sigma(\xi + \pi)\overline{c(\xi + \pi)} = 0 \quad \text{a.e.}$$

Multiplying with $\overline{\tilde{m}_\sigma(\xi)}$, and adding the two equations ($\sigma = 0, 1$) leads to (again, almost everywhere)

$$\begin{aligned} 0 = \overline{c(\xi)} &\left[m_0(\xi)\overline{\tilde{m}_0(\xi)} + \overline{\tilde{m}_0(\xi + \pi)}m_0(\xi + \pi)\right] \\ &+ \overline{c(\xi + \pi)}\left[m_0(\xi + \pi)\overline{\tilde{m}_0(\xi)} - \overline{\tilde{m}_0(\xi)}m_0(\xi + \pi)\right] = \overline{c(\xi)}, \end{aligned}$$

which proves that $u = 0$.

4. Next we derive an upper bound on $\sum_k \left[a_k f_k^0 + b_k f_k^1\right]$. With the notation $F(\xi) = \sum_k |\hat{f}(\xi + 2\pi k)|^2$, we have

$$\begin{aligned} &\left\|\sum_k \left[a_k f_k^0 + b_k f_k^1\right]\right\|^2 = 2\int_{-\infty}^{\infty} d\xi |\hat{f}(\xi)|^2 \left|\sum_k a_k e^{-2ik\xi} m_0(\xi) + \sum_k b_k e^{-2ik\xi} m_1(\xi)\right|^2 \\ &\le 2\int_0^\pi d\xi \left[|a(2\xi)|^2 \left(F(\xi)|m_0(\xi)|^2 + F(\xi + \pi)|m_0(\xi + \pi)|^2\right)\right. \\ &\qquad + |b(2\xi)|^2 \left(F(\xi)|\tilde{m}_0(\xi + \pi)|^2 + F(\xi + \pi)|\tilde{m}_0(\xi)|^2\right) \\ &\qquad \left. + 2|a(2\xi)||b(2\xi)||F(\xi)m_0(\xi)\tilde{m}_0(\xi + \pi) - F(\xi + \pi)m_0(\xi + \pi)\tilde{m}_0(\xi)|\right] \\ &\le 2\int_0^\pi d\xi \left[|a(2\xi)|^2 \left(\frac{B}{2\pi}M_0 + \frac{F(\xi) + F(\xi + \pi)}{2}|m_0(\xi)\tilde{m}_0(\xi + \pi) - m_0(\xi + \pi)\tilde{m}_0(\xi)|\right.\right. \\ &\qquad\qquad \left. + \left|\frac{F(\xi) - F(\xi + \pi)}{2}\right| (M_0\tilde{M}_0)^{1/2}\right) \\ &\qquad + |b(2\xi)|^2 \left(\frac{B}{2\pi}\tilde{M}_0 + \frac{F(\xi) + F(\xi + \pi)}{2}|m_0(\xi)\tilde{m}_0(\xi + \pi) - m_0(\xi + \pi)\tilde{m}_0(\xi)|\right. \\ &\qquad\qquad \left.\left. + \left|\frac{F(\xi) - F(\xi + \pi)}{2}\right| (M_0\tilde{M}_0)^{1/2}\right)\right] \\ &\le \left[B(M_0 + \Delta_0) + \frac{B - A}{2}(M_0\tilde{M}_0)^{1/2}\right] \sum |a_k|^2 \\ &\quad + \left[B(\tilde{M}_0 + \Delta_0) + \frac{B - A}{2}(M_0\tilde{M}_0)^{1/2}\right] \sum |b_k|^2, \end{aligned}$$

which implies the upper bound in (4.2).

5. To derive the lower bound in (4.2), we introduce a dual family $\tilde{f}_k^0$, $\tilde{f}_k^1$, defined by $\tilde{f}_k^\sigma(x) = \frac{1}{\sqrt{2}}\tilde{f}^\sigma\left(\frac{x}{2} - k\right)$, with

$$\widehat{\tilde{f}}^{\,\sigma}(\xi) = \tilde{m}_\sigma(\xi/2)\widehat{\tilde{f}}\,(\xi/2),$$

where $\tilde{m}_0(\xi)$ is as above, and $\tilde{m}_1(\xi) = e^{-i\xi}\overline{m_0(\xi + \pi)}$. One easily proves $\langle f_\ell^\sigma, \tilde{f}_m^\tau\rangle = \delta_{\sigma,\tau}\delta_{\ell,m}$. For $\sigma = \tau = 0$, for instance,

$$\langle f_\ell^0, \tilde{f}_m^0\rangle = 2\int d\xi m_0(\xi)\overline{\tilde{m}_0(\xi)}\hat{f}(\xi)\widehat{\tilde{f}}\,(\xi)e^{2i(\ell - m)\xi}$$

$$\begin{aligned}
&= \pi^{-1}\int_0^{2\pi} d\xi m_0(\xi)\overline{\tilde{m}_0(\xi)} e^{2i(\ell-m)\xi} \\
&= \pi^{-1}\int_0^{\pi} d\xi \left[m_0(\xi)\overline{\tilde{m}_0(\xi)} + m_0(\xi+\pi)\overline{\tilde{m}_0(\xi+\pi)}\right] e^{2i(\ell-m)\xi} \\
&= \delta_{\ell,m}.
\end{aligned}$$

6. The same arguments as in point 4, together with (4.4), prove that

$$\begin{aligned}
&\left\|\sum_k a_k\tilde{f}_k^0 + b_k\tilde{f}_k^1\right\|^2 \\
&\quad\le \left[A^{-1}(\max(M_0,\tilde{M}_0)+\Delta_0) + \frac{A^{-1}-B^{-1}}{2}(M_0\tilde{M}_0)^{1/2}\right]\sum_k\left[|a_k|^2+|b_k|^2\right] \\
&\quad= A'^{-1}\sum_k\left[|a_k|^2+|b_k|^2\right].
\end{aligned}$$

The lower bound in (4.2) now follows from a simple duality argument.

$$\begin{aligned}
\sum_k\left[|a_k|^2+|b_k|^2\right] &= \left\langle \sum_k\left[a_kf_k^0+b_kf_k^1\right], \sum_k\left[a_k\tilde{f}_k^0+b_k\tilde{f}_k^1\right]\right\rangle \\
&\le \left\|\sum_k\left[a_kf_k^0+b_kf_k^1\right]\right\| A'^{-1/2}\left(\sum_k\left[|a_k|^2+|b_k|^2\right]\right)^{1/2}. \qquad \square
\end{aligned}$$

An immediate corollary is the following.

COROLLARY 4.2. *We assume the same as in Lemma* 4.1. *If* $u = u_0 + u_1$ *is the unique decomposition of* $u \in E$ *into* $u_\sigma = \sum_k c_k^\sigma f_k^\sigma$, *then*

$$\frac{A'}{B'}\left[\|u_0\|^2+\|u_1\|^2\right] \le \|u\|^2 \le \frac{B'}{A'}\left[\|u_0\|^2+\|u_1\|^2\right].$$

Proof. By the same argument as for Lemma 3.1. □

5. Biorthogonal wavelet packets. We can now apply the biorthogonal splitting trick to the spaces V_0, W_j in the nonorthogonal decomposition (3.5). We start by choosing, among V_0 and all the W_j, an arbitrary subset of spaces to be split, and we apply the biorthogonal splitting trick to all of them. We end up with a different decomposition, in which all the "split" spaces are replaced by their two offspring. We can then repeat the procedure: choose an arbitrary subset, and split again. Every splitting corresponds to a replacement of the basis vectors as well. If, after L splitting steps, the subspace W_j has undergone $J \le L$ splittings, then the $\psi_{j,k}(x) = 2^{j/2}\psi(2^jx-k)$, $k \in \mathbb{Z}$, will have been replaced by $\psi^J_{j;\epsilon_1,\dots,\epsilon_k;\ell(x)} = 2^{(j-J)/2}\psi^J_{\epsilon_1,\dots,\epsilon_J}(2^{j-J}x-\ell)$, $\epsilon_n = 0$ or 1, $\ell \in \mathbb{Z}$, with $\hat{\psi}^J_{\epsilon_1,\dots,\epsilon_J}(\xi) = m_{\epsilon_1}(\xi/2),\dots,m_{\epsilon_J}(2^{-J}\xi)\,\hat{\psi}(2^{-J}\xi)$. The following theorem tells us that as long as we confine ourselves to a finite number of splitting steps, the result is still a Riesz basis.

THEOREM 5.1. *Suppose we start from the decomposition* (3.5), *with*

$$
\begin{aligned}
A\left[\sum_k |\alpha_k|^2 + \sum_{j,k} |\beta_{j,k}|^2\right] &\le \left\| \sum_{k\in\mathbb{Z}} \alpha_k \phi_{0,k} + \sum_{j=0}^{\infty}\sum_{k\in\mathbb{Z}} \beta_{j,k}\psi_{j,k} \right\|^2 \\
&\le B\left[\sum_k |\alpha_k|^2 + \sum_{j,k} |\beta_{j,k}|^2\right] .
\end{aligned}
\tag{5.1}
$$

Let us denote by Ψ^L_λ *the vectors obtained after* L *splitting steps, as described above (the label* λ *stands for* $J, j, \epsilon_1, \dots, \epsilon_K$ *and* ℓ*). Then the* Ψ^L_λ *still constitute a Riesz basis, and*

$$
A_L \sum_\lambda |\gamma_\lambda|^2 \le \left\| \sum_\lambda \gamma_\lambda \Psi^L_\lambda \right\|^2 \le B_L \sum_\lambda |\gamma_\lambda|^2 .
\tag{5.2}
$$

The constants B_L, A_L *are defined recursively by*

$$
B_0 = B, A_0 = A, \beta_0 = B/A, \alpha_0 = A/B,
$$

and

$$
\begin{aligned}
B_L &= \beta_{L-1}\left[B_{L-1}\mu_0 + (B_{L-1} - A_{L-1})\,\nu_0\right] \\
A_L &= \alpha_{L-1}\left[A^{-1}_{L-1}\mu_0 + \left(A^{-1}_{L-1} - B^{-1}_{L-1}\right)\nu_0\right]^{-1}
\end{aligned}
\tag{5.3}
$$

with

$$
\begin{aligned}
\beta_L &= \beta_{L-1}\left[B_{L-1}\mu_0 + (B_{L-1} - A_{L-1})\,\nu_0\right]\left[A^{-1}_{L-1}\mu_0 + \left(A^{-1}_{L-1} - B^{-1}_{L-1}\right)\nu_0\right] \\
\alpha_L &= \alpha_{L-1}\left[A^{-1}_{L-1}\mu_0 + \left(A^{-1}_{L-1} - B^{-1}_{L-1}\right)\nu_0\right]^{-1}\left[B_{L-1}\mu_0 + (B_{L-1} - A_{L-1})\,\nu_0\right]^{-1}
\end{aligned}
\tag{5.4}
$$

and

$$
\begin{aligned}
\mu_0 &= \max(M_0, \tilde{M}_0) + \Delta_0 \\
\nu_0 &= \tfrac{1}{2}(M_0\tilde{M}_0)^{1/2}
\end{aligned}
$$

($M_0, \tilde{M}_0, \Delta_0$ *as defined in Lemma* 4.1).

Proof.

1. We will work by induction on L, the number of splittings. Suppose that we have gone through ℓ splitting steps, resulting in a (nonorthogonal) decomposition of $L^2(\mathbb{R})$, i.e.,

$$
L^2(\mathbb{R}) = \bigoplus_m E_{\ell,m};
$$

in each $E_{\ell,m}$ we have a Riesz basis $F_{\ell,m;k}$, $k \in \mathbb{Z}$. Assume that

$$
A_\ell \sum_{m,k} |c_{m,k}|^2 \le \left\| \sum_{m,k} c_{m,k} F_{\ell,m;k} \right\|^2 \le B_\ell \sum_{m,k} |c_{m,k}|^2
$$

and that for arbitrary $u \in L^2(\mathbb{R})$, $u = \sum_m u_{\ell,m}$, with $u_{\ell,m} \in E_{\ell,m}$, we have

$$\alpha_\ell \sum_m \|u_{\ell,m}\|^2 \leq \|u\|^2 \leq \beta_\ell \sum_m \|u_{\ell,m}\|^2. \tag{5.5}$$

We now choose an arbitrary subset of the $\{E_{\ell,m}; m \in \mathbb{Z}\}$ of spaces to be split, and we apply the biorthogonal splitting trick to each of them. If $E_{\ell,n}$ is a space that gets split into $E^0_{\ell,n} \oplus E^1_{\ell,n}$, then for arbitrary

$$u_{\ell,n} = \sum_k c^0_{n,k} F^0_{\ell,n;k} + \sum_k c^1_{n,k} F^1_{\ell,n;k} = u^0_{\ell,n} + u^1_{\ell,n},$$

applying Lemma 4.1 leads to

$$\begin{aligned}&\left[A_\ell^{-1}\mu_0 + \left(A_\ell^{-1} - B_\ell^{-1}\right)\nu_0\right]^{-1} \sum_{\sigma=0,1} \sum_k |c^\sigma_{n,k}|^2 \\ &\leq \|u_{\ell,n}\|^2 \leq \left[B_\ell\mu_0 + (B_\ell - A_\ell)\nu_0\right] \sum_{\sigma=0,1} \sum_k |c^\sigma_{n,k}|^2.\end{aligned}$$

Bringing all these inequalities together, combining with (5.5), and relabeling the $F^\sigma_{\ell,n;k}$ as $F_{\ell+1,m;k'}$, we obtain

$$\begin{aligned}&\alpha_\ell \left[A_\ell^{-1}\mu_0 + \left(A_\ell^{-1} - B_\ell^{-1}\right)\nu_0\right]^{-1} \sum_k |c_{m,k}|^2 \leq \left\|\sum_{m,k} c_{m,k} F_{\ell+1,m;k}\right\|^2 \\ &\leq \beta_\ell \left[B_\ell\mu_0 + (B_\ell - A_\ell)\nu_0\right] \sum_k |c_{m,k}|^2.\end{aligned} \tag{5.6}$$

2. From Corollary 4.2 and (5.5) we also obtain

$$\alpha_{\ell+1} \sum_m \|u_{\ell+1,m}\|^2 \leq \|u\|^2 \leq \beta_{\ell+1} \sum_m \|u_{\ell+1,m}\|^2, \tag{5.7}$$

with

$$\begin{aligned}\beta_{\ell+1} &= \beta_\ell \left[B_\ell\mu_0 + (B_\ell - A_\ell)\nu_0\right] \left[A_\ell^{-1}\mu_0 + \left(A_\ell^{-1} - B_\ell^{-1}\right)\nu_0\right], \\ \alpha_{\ell+1} &= \alpha_\ell \left[B_\ell\mu_0 + (B_\ell - A_\ell)\nu_0\right]^{-1} \left[A_\ell^{-1}\mu_0 + \left(A_\ell^{-1} - B_\ell^{-1}\right)\nu_0\right]^{-1}.\end{aligned}$$

(5.6) and (5.7) can be used for the next induction step.

3. To start it all (at $L = 0$), we need (5.1), together with Lemma 3.1, which leads to

$$\begin{aligned}A_0 &= A, & B_0 &= B \\ \alpha_0 &= A/B, & B_0 &= B/A.\end{aligned} \qquad \square$$

Remarks.

1. If $A = B = 1$ and $m_0 = \tilde{m}_0$, then $M_0 = \tilde{M}_0 = 1$, $\Delta_0 = 0$, and $A_L = B_L = 1$ for every L; in the orthonormal case we recover the exact estimates for orthonormal wavelet packets.

2. In the nonorthonormal case, B_L and A_L^{-1} increase very rapidly with L: one easily checks that (5.2) and (5.3) imply that $B_L \sim CM^{2^L}$ for large L. In Chui and Li [2] a different technique is used to derive bounds similar to (5.2), with $\log B_L \sim$

const. L, i.e., only exponential growth for B_L. The bounds of Chui and Li do not seem to collapse to the optimal bounds 1 in the orthonormal case, however. They also restrict themselves to the case where one chooses, at every splitting step, to split every available subspace; this is probably not a crucial restriction. The estimates in the next section show that B_L grows at least exponentially with L.

3. For every possible choice of splittings, the dual basis of the resulting Riesz basis can be constructed by applying exactly the same choice of splittings on the original dual decomposition (with the $\tilde{\psi}_{j,k}$ instead of the $\psi_{j,k}$), and using $\tilde{m}_0, \tilde{m}_1$ instead of m_0, m_1 at every splitting step. This follows from the constructions in §4; see also Chui and Li [2].

6. Instability of arbitrary biorthogonal wavelet packets. If we choose to restrict to at most L splittings, then the previous section tells us that we will still have a Riesz basis, even though the constants involved may be large. In the orthonormal case, a very beautiful special wavelet packet basis resulted from splitting every W_j exactly j times. In this decomposition, the total number of splitting steps is not limited, and Theorem 5.1 does not guarantee that the biorthogonal analog leads to a Riesz basis. We shall show in this section that in fact we don't have a Riesz basis (except in the orthonormal case). Define, as in the orthonormal case,

$$\hat{\psi}_\epsilon(\xi) = \left[\prod_{j=1}^{N} m_{\epsilon_j}(2^{-j}\xi)\right] \hat{\phi}(2^{-N}\xi),$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_N)$ is a sequence of length N ($|\epsilon| = N$) consisting of only zeros and ones. We start by proving several lemmas about the $\hat{\psi}_\epsilon$.

LEMMA 6.1. *There exists a constant $C > 0$ such that*

$$\int d\xi |\hat{\psi}_\epsilon(\xi)|^2 \geq C \int_{|\xi| \leq 2^N \pi} d\xi \prod_{j=1}^{N} |m_{\epsilon_j}(2^{-j}\xi)|^2 \tag{6.1}$$

for all N and ϵ with $|\epsilon| = N$.

Proof.

1. Remember (see §3) that there exists a compact set K, congruent with $[-\pi, \pi]$ modulo 2π, so that $|\hat{\phi}(\xi)| \geq C^2 > 0$ for all $\xi \in K$.

2. We have

$$|\hat{\psi}_\epsilon(\xi)| = \prod_{j=1}^{N} |m_{\epsilon_j}(2^{-j}\xi)| |\hat{\phi}(2^{-N}\xi)|,$$

hence

$$\begin{aligned}\int d\xi |\hat{\psi}_\epsilon(\xi)|^2 &\geq \int_{\xi \in 2^N K} d\xi \prod_{j=1}^{N} |m_{\epsilon_j}(2^{-j}\xi)|^2 |\hat{\phi}(2^{-N}\xi)|^2 \\ &\geq C \int_{\xi \in 2^N K} d\xi \prod_{j=1}^{N} |m_{\epsilon_j}(2^{-j}\xi)|^2 \\ &= C \int_{|\xi| \leq 2^N \pi} d\xi \prod_{j=1}^{N} |m_{\epsilon_j}(2^{-j}\xi)|^2,\end{aligned}$$

where we have used the congruency of K with $[-\pi, \pi]$, and the 2π-periodicity of the m_{ϵ_j}. □

LEMMA 6.2. *Define $p(\xi) = |m_0(\xi)|^2 + |m_1(\xi)|^2$. Then for some $C > 0$,*

$$\sum_{\epsilon \in S_N} \|\psi_\epsilon\|^2 \geq C2^N \int_{|\xi| \leq \pi} d\xi \prod_{j=0}^{N-1} p(2^j\xi), \tag{6.2}$$

where S_N is the set of all sequences $\epsilon = (\epsilon_1, \dots, \epsilon_N)$ of length N and consisting of only zeros and ones.

Proof. The proof follows immediately from (6.1) by summing over the 2^N sequences ϵ with length N, and by changing the integration variable. □

The following lemma will allow us to compute a lower bound for the right-hand side of (6.2).

LEMMA 6.3. *The function $p(\xi) = |m_0(\xi)|^2 + |m_1(\xi)|^2$ satisfies*

$$p(\xi)p(\xi + \pi) \geq 1.$$

Moreover, if $m_0 \neq \tilde{m}_0$ (nonorthogonal case), then

$$p(\xi)p(\xi + \pi) > 1 \quad a.e.$$

Proof.

1. We know (see §4) that

$$m_0(\xi)\overline{\tilde{m}_0(\xi)} + m_0(\xi + \pi)\overline{\tilde{m}_0(\xi + \pi)} = 1. \tag{6.3}$$

By Cauchy–Schwarz, this implies

$$\left[|m_0(\xi)|^2 + |\tilde{m}_0(\xi + \pi)|^2\right] \left[|\tilde{m}_0(\xi)|^2 + |m_0(\xi + \pi)|^2\right] \geq 1 \tag{6.4}$$

or

$$p(\xi)p(\xi + \pi) \geq 1 \left(\text{use } m_\sigma(\xi) = e^{-i\xi}\overline{\tilde{m}_{1-\sigma}(\xi + \pi)}\right).$$

2. Equality in (6.4) is only possible for those ξ for which

$$\tilde{m}_0(\xi) = \alpha(\xi)m_0(\xi), \qquad \overline{m_0(\xi + \pi)} = \alpha(\xi)\overline{\tilde{m}_0(\xi + \pi)}$$

for some $\alpha(\xi)$. For such ξ,

$$\tilde{m}_0(\xi)\overline{\tilde{m}_0(\xi + \pi)} - \overline{m_0(\xi + \pi)}m_0(\xi) = 0. \tag{6.5}$$

3. Suppose that (6.5) were true for all ξ. If we extend the trigonometric polynomials from $z = e^{-i\xi}$ on the torus to all $z \in \mathbb{C}$ (extending $M_0(e^{-i\xi}) = m_0(\xi)$), then the identity (6.5) would still hold for all z,

$$\tilde{M}_0(z)\bar{\tilde{M}}(-z^{-1}) - M_0(z)\bar{M}_0(-z^{-1}) = 0, \tag{6.6}$$

where for $A(z) = \sum_n a_n z^n$, we use the notation $\bar{A}(z) = \sum_n \bar{a}_n z^n$. On the other hand, extension of (6.3) gives

$$M_0(z)\bar{\tilde{M}}_0(z^{-1}) + M_0(-z)\bar{\tilde{M}}_0(-z^{-1}) = 1,$$

which means that $M_0(z)$ and $\bar{\tilde{M}}(-z^{-1})$ share no zeros. It then follows from (6.6) that $\tilde{M}_0(z)$ is zero whenever $M_0(z)$ is. Similarly one concludes that $M_0(z)$ is zero whenever $\tilde{M}_0(z)$ is. Since both are polynomials (up to multiplication by an integer power of z), $M_0 \equiv \tilde{M}_0$ follows.

4. If $\tilde{m}_0 \neq m_0$, one finds therefore that the left-hand side of (6.5) is a nontrivial trigonometric polynomial. It follows that (6.5) can only hold in a finite number of ξ_j. Consequently, (6.4) is a strict inequality except in this finite number of ξ_j, which implies that $p(\xi)p(\xi+\pi) > 1$ almost everywhere. □

We are now ready for the following theorem.

THEOREM 6.4. *There exist $C > 0$, $\lambda > 1$ so that, for all $N \in \mathbb{N}$, $N \geq 1$,*

$$\sum_{\substack{\epsilon \\ |\epsilon|=N}} \|\psi_\epsilon\|^2 \geq C 2^N \lambda^N. \tag{6.7}$$

Proof.

1. By Lemma 6.2, we only need to prove that

$$\int_{-\pi}^{\pi} d\xi \prod_{j=0}^{N-1} p(2^j\xi) \geq C\lambda^N.$$

2. By Jensen's inequality,

$$\log\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} d\xi \prod_{j=0}^{N-1} p(2^j\xi)\right] \geq \frac{1}{2\pi}\int_{-\pi}^{\pi} d\xi \log\left[\prod_{j=0}^{N-1} p(2^j\xi)\right]$$

$$= \frac{1}{2\pi}\sum_{j=0}^{N-1}\int_{-\pi}^{\pi} d\xi \log p(2^j\xi) = \frac{N}{2\pi}\int_{-\pi}^{\pi} d\xi \log p(\xi).$$

3. Since $p(\xi)p(\xi+\pi) > 1$ a.e., it follows that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} d\xi \log p(\xi) = \frac{1}{2\pi}\int_{0}^{\pi} d\xi \log[p(\xi)p(\xi+\pi)] = \gamma > 0. \qquad \square$$

Consequently,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} d\xi \prod_{j=0}^{N-1} p(2^j\xi) \geq e^{N\gamma}.$$

Remark. The argument in this proof is borrowed from the proof of Theorem 3 in Coifman, Meyer, and Wickerhauser [10].

This suffices to prove the instability claimed above. If the collection

$$\mathcal{C} = \{\phi(\cdot - k); k \in \mathbb{Z}\} \cup \{\psi_\epsilon(\cdot - k), k \in \mathbb{Z} \text{ and } |\epsilon| = N, \epsilon_N = 1, N \in \mathbb{N}\backslash\{0\}\}$$

were a Riesz basis for $L^2(\mathbb{R})$, then it would follow that the L^2-norms of all these functions could be bounded uniformly by some constant C. (A Riesz basis is the image of an orthonormal basis under a continuous map.) In particular it would follow that, for all $N \in \mathbb{N}$,

$$\sum_{\epsilon \in S_N} \|\psi_\epsilon\|^2 \leq C\#S_N = C2^N.$$

This is contradicted by (6.7); the collection $\mathcal{C}$ does therefore not constitute a Riesz basis.

REFERENCES

[1] C. K. Chui (1992), *On cardinal spline wavelets*, in Wavelets and Their Applications, M. B. Ruskai et al., eds., Jones and Bartlett, Boston, pp. 419–438.

[2] C. K. Chui and C. Li (1993), *Nonorthogonal wavelet packets*, SIAM J. Math. Anal., 24, pp. 712–738.

[3] C. K. Chui and J. Z. Wang (1992), *A cardinal spline approach to wavelets*, Proc. Amer. Math. Soc. 113, pp. 785–793; *On compactly supported spline wavelets and a duality principle*, Trans. Amer. Math. Soc., 330, pp. 903–915.

[4] A. Cohen (1990a), *Ondelettes, analyses multirésolutions et filtres miroir en quadrature*, Ann. Inst. H. Poincaré Anal. Non Linéaire 7, pp. 439–459.

[5] ———(1990b), *Ondelettes, analyses multirésolutions et traitement numérique du signal*, Ph. D. thesis, Université Paris—Dauphine.

[6] A. Cohen and Q. Sun (1993), *On the necessary and sufficient condition for generating an orthonormal wavelet basis*, submitted.

[7] A. Cohen and I. Daubechies (1992), *A stability criterion for biorthogonal wavelet bases and their related subband coding schemes*, Duke Math. J., 68, pp. 313–335.

[8] ——— (1993), *Orthonormal bases of compactly supported wavelets* III: *Better frequency localization*, SIAM J. Math. Anal., 24, pp. 520–527.

[9] A. Cohen, I. Daubechies, and J. C. Feauveau (1992), *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45, pp. 485–560.

[10] R. Coifman, Y. Meyer, and M. V. Wickerhauser (1992), *Size properties of wavelet packets*, in Wavelets and Their Applications, M. B. Ruskai et al., eds., Jones and Bartlett, Boston, pp. 453–470.

[11] R. Coifman and M. V. Wickerhauser (1992), *Entropy-based algorithms for best basis selection*, IEEE Trans. Inform. Theory 38, pp. 713–718.

[12] I. Daubechies (1988), *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41, pp. 909–996.

[13] W. Lawton (1991), *Necessary and sufficient conditions for constructing orthonormal wavelet bases*, J. Math. Phys. 32, pp. 57–61.

[14] S. Mallat (1989), *Multiresolution approximation and wavelets*, Trans. Amer. Math. Soc., 315, pp. 69–88.

[15] Y. Meyer (1990), *Ondelettes ét opérateurs,* I: *Ondelettes,* II: *Opérateurs de Calderón-Zygmund,* III: *Opérateurs multilinéaires*, Hermann, Paris. (An English translation by the Cambridge University Press will appear in 1993.)

[16] M. B. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, and L. Raphael, eds. (1992), *Wavelets and Their Applications*, Jones and Bartlett, Boston.

SIAM J. MATH. ANAL.
Vol. 24, No. 5, pp. 1355–1360, September 1993

014

AN ARITHMETIC CHARACTERIZATION OF THE CONJUGATE QUADRATURE FILTERS ASSOCIATED TO ORTHONORMAL WAVELET BASES*

ALBERT COHEN[†] AND QIYU SUN[‡]

Abstract. Let $h = \{h_n\}$ be a sequence of complex numbers with finite length such that $H(\omega) = \sum h_n \exp(-2\pi in\omega)$ satisfies the identity $|H(\omega)|^2 + |H(\omega + \frac{1}{2})|^2 = 1$ and $H(0) = 1$, i.e., h is the impulse response of a conjugate quadrature filter. In this paper, we give a characterization, by the real roots of $H(\omega)$, of the sequences h that generate an orthonormal wavelet basis in the sense of the theory developed by Meyer and Daubechies. This result leads to a counterexample to Pollen's conjecture.

Key words. wavelet bases, conjugate quadrature filters, Pollen's conjecture

AMS) subject classification. 42C05

1. Introduction and results.

A conjugate quadrature filter (CQF) is a sequence $h = \{h_n\}_{n\in Z}$ of complex numbers with finite length, i.e., there exists a positive integer N such that $h_n = 0$ for $|n| > N$ that satisfies

$$2\sum_m h_m \overline{h}_{m+2n} = \delta(n) \quad \text{for all } n, \tag{1}$$

and

$$\sum_n h_n = 1, \tag{2}$$

or equivalently, whose Fourier transform defined by

$$H(\omega) = \sum h_n \exp(-2\pi in\omega) \tag{3}$$

satisfies

$$|H(\omega)|^2 + |H(\omega + \tfrac{1}{2})|^2 = 1 \tag{4}$$

and

$$H(0) = 1. \tag{5}$$

Here $\delta(m) = 1$ for $m = 0$ and $\delta(m) = 0$ otherwise. Let V denote the set of sequences h that satisfy (1) and (2). For every h in V, following the theory developed in [3], [6], and [7], we construct two functions φ and ψ by

$$\hat{\varphi}(\omega) = \prod_{k=1}^{\infty} H\left(\frac{\omega}{2^k}\right) \tag{6}$$

and

$$\hat{\psi}(\omega) = \exp(-2\pi i\omega)H\left(\omega + \frac{1}{2}\right)\hat{\varphi}\left(\frac{\omega}{2}\right), \tag{7}$$

*Received by the editors August 7, 1991; accepted for publication (in revised form) January 27, 1993.

[†] CEREMADE, Université de Paris Dauphine, Place de Lattre de Tassigny, 75775 Paris Codex 16, France.

[‡] Center for Mathematical Sciences, Zhejiang University, Hangzhou, Zhejiang 310027, People's Republic of China. This author is partially supported by Zhejiang Provincial Natural Science Foundation of China.

which represent, respectively, the Fourier transform of a scaling function ϕ and a wavelet function ψ in the sense of tempered distributions. We used the normalization $\hat{f}(\omega) = \int f(x)e^{2i\pi\omega x}dx$ here.

Lawton proved that for h in V, $\{\psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x - k)\}_{j,k\in\mathbb{Z}}$ forms a tight frame of $L^2(\mathbb{R})$ [4]. It is known that there exists h in V such that $\{\psi_{j,k}\}_{j,k\in\mathbb{Z}}$ does not form an orthonormal basis of $L^2(\mathbb{R})$. In fact, $\{\psi_{j,k}\}$ is an orthonormal basis if and only if $\{\varphi(x-k)\}_{k\in\mathbb{Z}}$ is orthonomal. By Poisson's summation formula, this is equivalent to

$$(8) \qquad \sum_k |\hat{\varphi}(\omega - k)|^2 = 1 \quad \text{for all } \omega \in \mathbb{R}.$$

We say that h generates an orthonormal wavelet basis if $\{\psi_{j,k}\}_{j,k\in\mathbb{Z}}$ is of that type. Hence to prove that h generates an orthonormal basis it suffices to prove that (8) holds for $\hat{\varphi}$ defined by the infinite product formula (6).

Mallat [6] proved, using the Lebesgue's dominated convergence theorem, that if $H(\omega) \neq 0$ for $|\omega| \leqq \frac{1}{4}$, then (8) holds for all ω in $\mathbb{R}$. A compact subset K of $\mathbb{R}$ is said to be congruous to $\left[-\frac{1}{2}, \frac{1}{2}\right]$ if and only if mes $(K) = 1$ and for almost every ω in $\left[-\frac{1}{2}, \frac{1}{2}\right]$, there exists a β in K such that $\omega - \beta$ is an integer.

Cohen [2] extended Mallat's result and proved that (8) holds if and only if there exists a compact subset K of $\mathbb{R}$ congruous to $\left[-\frac{1}{2}, \frac{1}{2}\right]$ and containing zero in its interior such that $H(2^{-j}\omega) \neq 0$ for all $\omega \in K$ and positive integers j. This result can be expressed in the following way: the identity (8) holds if and only if

$$(9) \qquad \sum_k |\hat{\varphi}(\omega - k)|^2 \neq 0 \quad \text{for all } \omega \in \mathbb{R}.$$

We shall use this important property in the proof of the main theorem of this paper.

Lawton [4], [5] uses an eigenvector relationship between a linear operator on $l^2(\mathbb{Z})$, the set of square summable sequences, to characterize sequences $h \in V$ that generate orthonormal wavelet bases.

Let $R(H)$ be the set of real roots of H in $(0, 1)$. In this paper, we shall characterize these sequences by an arithmetic condition on $R(H)$.

For relatively prime positive odd integers m and n, let $\mathrm{Ind}_2\, n$ be the least positive integer k such that $2^k = 1 \pmod{n}$ and define the number $m_i = 2^i m + n \pmod{2n}$ for $l \leqq i \leqq \mathrm{Ind}_2\, n$. In the next section, we shall prove the following result.

THEOREM 1. *Let $h \in V$. Then the following propositions are equivalent*:

(i) *h generates an orthonormal wavelet basis*;

(ii) *There exist no relatively prime positive odd numbers m and $n \neq 1$ such that $H(m_i/2n) = 0$ for all $1 \leqq i \leqq \mathrm{Ind}_2$ n*;

(iii) *For all ω in* $]0, 1[$, $\lim_{N\to+\infty} \prod_{k=0}^{N} H(2^k\omega) = 0$.

Remarks.

• The proof of Theorem 1 will use the property that the zeros of $H(\omega)$ are isolated points. This result can thus be extended to the case where h is an infinite sequence with exponential decay.

• Pollen [8] conjectured that if h does not generate an orthonormal basis, then there exists an odd integer $n \geqq 3$ and a sequence $\tilde{h}$ in V such that $H(\omega) = \tilde{H}(n\omega)$. Thus, $H(\omega)$ should vanish at all the points $\{m/2n | m \text{ is odd}\} = R_n$. We shall construct counterexamples to this conjecture by taking two relatively prime odd integers m and n, $m \geqq 1$ and $n \geqq 3$ such that the set $\{m_i/2n, 1 - (m_i/2n) | 1 \leqq i \leqq \mathrm{Ind}_2\, n\} \cup \{\frac{1}{2}\}$ is strictly included into R_n. We will then use the following constructive result.

THEOREM 2. *Let m and $n \neq 1$ be two relatively prime positive odd integers. Then there exists an $h \in V$ such that h does not generate an orthonormal basis and $R(H) = \{m_i/2n, 1-(m_i/2n), \frac{1}{2} | 1 \leqq i \leqq \operatorname{Ind}_2 n\}$.*

Observe that $H(\omega) = 0$ implies $H(1-\omega) = 0$ for a sequence $h \in V$ of real numbers. Both of the authors obtained the results above independently at about the same time.

2. Proof of Theorem 1. (a) We start by the assumption that h does not generate an orthonormal wavelet basis and show that $H(\omega)$ vanishes on a set of points of the type $\{m_i/2n | 1 \leqq i \leqq \operatorname{Ind}_2 n\}$.

From (9), we know that for a certain ω_0 in $[0,1]$, $\hat{\varphi}(\omega_0 + l) = 0$ for all l in $\mathbb{Z}$. Consequently, for all l in $\mathbb{Z}$, we can define

$$j(l) = \min\{j > 0 | H(2^{-j}(\omega_0 + l)) = 0\} \tag{10}$$

and

$$\nu(l) = 2^{-j(l)}(\omega_0 + l)(\text{mod } 1), \tag{11}$$

the associated root of $H(\omega)$. Because $H(\omega)$ has a finite number of roots in $[0,1]$, the family $\{\nu(l)\}_{l\in\mathbf{Z}}$ contains only a finite number of points. We shall proceed in four steps to prove that this family contains the announced type of set.

Step 1. There exists a sequence $\{l_k\}_{k \geq 0}$ such that $j(l_{k+1}) > j(l_k)$.

We can construct this sequence recursively: take $l_0 = 0$ and suppose that we have defined the sequence up to l_k. We then write

$$n_k = j(l_k) \tag{12}$$

and

$$\mu_k = \nu(l_k) = 2^{-n_k}(\omega_0 + l_k)(\text{mod } 1). \tag{13}$$

We choose for l_{k+1} the value

$$l_{k+1} = l_k + 2^{n_k - 1}. \tag{14}$$

To check that n_{k+1} is strictly superior to n_k, we remark that for all $n < n_k$,

$$\begin{aligned} H(2^{-n}(\omega_0 + l_{k+1})) &= H(2^{-n}(\omega_0 + l_k) + 2^{n_k - n - 1}) \\ &= H(2^{-n}(\omega_0 + l_k)) \neq 0 \end{aligned}$$

and

$$\begin{aligned} |H(2^{-n_k}(\omega_0 + l_{k+1}))| &= |H(2^{-n_k}(\omega_0 + l_k) + \tfrac{1}{2})| \\ &= |H(\mu_k + \tfrac{1}{2})| = 1 \neq 0. \end{aligned}$$

It follows that $n_{k+1} = j(l_{k+1})$ is strictly superior to $n_k = j(l_k)$, and the sequence $\{j(l)\}_{l\in\mathbf{Z}}$ contains an infinite number of elements. We shall reuse the sequence μ_k, which satisfies

$$\mu_{k+1} = 2^{n_k - n_{k+1}}(\mu_k + \tfrac{1}{2}). \tag{15}$$

Step 2. The family $\{\nu(l)\}_{l\in\mathbf{Z}}$ contains only rational numbers.

It suffices to prove that ω_0 is rational. Since there is an infinite number of $j(l)$ and only finitely many $\nu(l)$, it is possible to find l and l' such that $\nu(l) = \nu(l')$ and $j(l) \neq j(l')$. Consequently, $2^{-j(l)}(\omega_0+l) - 2^{-j(l')}(\omega_0+l')$ is an integer. It follows that ω_0 is a rational number and so are the $\{\nu(l)\}_{l\in\mathbf{Z}}$.

Step 3. $\nu(l)$ can be written in an irreducible form as $\nu(l) = p(l)/2n(l)$, where $p(l)$ and $n(l)$ are odd and $n(l) > 1$.

Let us prove that all the other possibilities are not coherent with the hypotheses.

If $\nu(l)$ was an integer or a half integer, since $\nu(l) = 2^{-j(l)}(\omega_0 + l)$ and $j(l) \geqq 1$, this would imply that ω_0 is an integer. This is impossible since we have

$$\sum_{k\in\mathbf{Z}} |\hat{\varphi}(k)|^2 = |\hat{\varphi}(0)|^2 = 1. \tag{16}$$

Now, let us suppose that the denominator of $\nu(l)$ is a multiple of 4, and let us construct the sequence $\{\mu_k\}_{k\geqq 0}$ as in step 1, with $\mu_0 = \nu(l)$.

We can write μ_0 in an irreducible form:

$$\mu_0 = 2^{-j_0}\frac{p_0}{n}, \tag{17}$$

where p_0 and n are odd and $j_0 \geqq 2$.

By (15), we then have

$$\mu_1 = 2^{n_0-n_1}\left(\mu_0 + \frac{1}{2}\right) = 2^{n_0-n_1-j_0}\left(\frac{p_0 + 2^{j_0-1}n}{n}\right) = 2^{-j_1}\left(\frac{p_1}{n}\right),$$

which is the irreducible form for μ_1 since $p_1 = p_0 + 2^{j_0-1}n$ is odd and prime together with n. It is clear that $j_1 > j_0$ and by iterating this process we have $\mu_k = 2^{-j_k}(p_k/n)$ with $j_k > j_{k-1}$. This clearly contradicts the hypothesis that the $\nu(l)$ are finitely many (modulo 1).

Finally, if $\nu(l)$ had an odd denominator, we could write $\mu_0 = p_0/n$, and this would lead to

$$\mu_1 = 2^{n_0-n_1}\left(\mu_0 + \frac{1}{2}\right) = 2^{n_0-n_1-1}\left(\frac{2p_0 + n}{n}\right) = 2^{-j_1}\left(\frac{p_1}{n}\right)$$

with $j_1 \geqq 2$, which is exactly the previous situation.

The only solution is $\nu(l) = p(l)/2n(l)$.

Step 4. Now we shall construct the set of zero of $H(\omega)$ that has been announced.

First remark that the sequence μ_k can also be generated by taking

$$\mu_{k+1} = 2^{n_k - n_{k+1}}\left(\mu_k - \tfrac{1}{2}\right). \tag{18}$$

For a given $\mu_0 = p_0/2n$ we have $\mu_0 - \frac{1}{2} = (p_0 - n)/2n$ and $\mu_0 + \frac{1}{2} = (p_0 + n)/2n$. Clearly, both $p_0 - n$ and $p_0 + n$ are even, but since their difference is $2n$, one of them cannot be a multiple of 4.

Consequently, to preserve the arithmetic structure imposed by the third step, we shall take

$$\mu_1 = \begin{cases} \dfrac{1}{2}\left(\mu_0 - \dfrac{1}{2}\right) & \text{if } \dfrac{p_0 - n}{2} \text{ is odd,} \\[2ex] \dfrac{1}{2}\left(\mu_0 + \dfrac{1}{2}\right) & \text{if } \dfrac{p_0 + n}{2} \text{ is odd.} \end{cases} \tag{19}$$

By iteration we obtain a sequence of numbers:

$$\mu_k = \frac{p_k}{2n} \quad \text{with } \{p_{k+1}\} = \left\{\frac{p_k+n}{2}, \frac{p_k-n}{2}\right\} \cap 2\mathbb{Z}+1.$$

Remark that by choosing μ_0 in $[-\frac{1}{2}, \frac{1}{2}]$ rather than in $[0, 1]$, the sequence μ_k will stay in $[-\frac{1}{2}, \frac{1}{2}]$. It is thus necessarily a cycle.

Let us now consider the shifted cyclic sequence

$$\gamma_k = \mu_k + \frac{1}{2} = \frac{p_k+n}{2n} = \frac{q_k}{n}. \tag{20}$$

By (19) we obtain

$$\gamma_k = 2\gamma_{k+1} (\text{mod } 1), \tag{21}$$

which shows that γ_k is a cycle for the transformation $\omega \mapsto 2\omega(\text{mod } 1)$. As a consequence, there exists an odd number $m < n$ such that, after reordering the cycle γ_k (in the reverse order), we have

$$\gamma_i = \frac{2^i m}{n} (\text{mod } 1), \quad 0 \leqq i \leqq \text{Ind}_2\, n. \tag{22}$$

It follows that $H(\omega)$ vanishes on the set of point $\{m_i/2n | 1 \leqq i \leqq \text{Ind}_2\, n\}$.

We have thus proved (ii) $\Rightarrow$ (i).

(b) We now prove (iii) $\Rightarrow$ (ii). Suppose that (ii) is not true and that $H(\omega)$ vanish on such a set. Then it follows that $|H(\gamma_i)| = 1$, where $\{\gamma_i\}$ is the cycle given by (22). Consequently, for all $N > 0$,

$$\left| \prod_{k=0}^{N} H(2^k \gamma_0) \right| = 1, \tag{23}$$

and $\gamma_0 = m/n$ is in $]0, 1[$, which contradicts (iii).

(c) Finally, we show that (i) $\Rightarrow$ (iii). Since φ is both compactly supported and square-integrable, it is in $L^1(\mathbb{R})$ by the Cauchy–Schwarz formula. It follows that $\hat{\varphi}(\omega)$ has to go to zero when $\omega \to +\infty$ or $-\infty$.

By (8), for all ω in $]0, 1[$, there exist an integer l such that $\hat{\varphi}(\omega + l) \neq 0$. Let us write $\omega_l = \omega + l \neq 0$.

We have

$$\begin{aligned} \hat{\varphi}(2^N \omega_1) &= \prod_{k=0}^{N-1} H(2^k \omega_l) \hat{\varphi}(\omega_l) \\ &= \prod_{k=0}^{N-1} H(2^k \omega) \hat{\varphi}(\omega_l), \end{aligned}$$

and thus $\prod_{k=0}^{N-1} H(2^k \omega)$ must go to zero when $N \to +\infty$, which implies (iii).

This concludes the proof of Theorem 1.

From this proof we see that the cancellation of $H(\omega)$ at the points $m_i/2n$ is a necessary and sufficient condition for the orthonormality to fail. (In fact one can prove that the numbers $\nu(l)$ are exactly the points $m_i/2n$ modulo 1.)

3. Proof of Theorem 2. Let us define

$$R_n^1 = \left\{\frac{m_i}{2n}, 1 - \frac{m_i}{2n} | 1 \leqq i \leqq \mathrm{Ind}_2\, n\right\}, \tag{24}$$

and let R_n^2 be the complement of $R_n^1 \cup \{\frac{1}{2}\}$ in R_n. Recall that $R_n = \{m/2n | m \text{ is odd}\}$ is the set of all roots of $1 + e^{2\pi i n\omega}$.

In order to construct a trigonometric polynomial $H(\omega)$ that satisfies (4) and (5) and vanished only on the set $R_n^1 \cup \{\frac{1}{2}\}$, we need to define

$$P(\omega) = \prod_{\omega_i \in R_n^1} |e^{4\pi i\omega} - e^{4\pi i\omega_i}|^2, \tag{25}$$

and for each ω_j in R_n^2,

$$Q_j(\omega) = \cos 2\pi\omega_j \cos 2\pi\omega \prod_{\omega_k \in R_n^2 - \{\omega_j, 1-\omega_j\}} |e^{4\pi i\omega} - e^{4\pi i\omega_k}|^2. \tag{26}$$

Consequently, $Q_j(\omega)$ vanishes on $R_n^2 = \{\omega_j, 1 - \omega_j\}$ and we have $Q_j(\omega_j) = Q_j(1 - \omega_j) > 0$. It follows that we can find some strictly positive numbers ε_j such that

$$R(\omega) = |(1 + e^{2\pi i n\omega})/2|^2 + \sin^2 2\pi\omega P(\omega) \sum \varepsilon_j Q_j(\omega) \tag{27}$$

is a nonnegative trigonometric polynomial. Clearly, $R(\omega)$ vanishes only on the set $R_n^1 \cup \{\frac{1}{2}\}$ and one checks easily that it satisfies $R(0) = 1$ and $R(\omega) + R(\omega + \frac{1}{2}) = 1$.

It suffices then to build a trigonometric polynomial H that satisfies $|H(\omega)|^2 = R(\omega)$, which exists by using the Riesz lemma. This concludes the proof of Theorem 2.

From this result we can construct counterexamples to Pollen's conjecture. We can take, for example,

$$R_{15}^1 \cup \left\{\frac{1}{2}\right\} = \left\{\frac{1}{30}, \frac{7}{30}, \frac{11}{30}, \frac{13}{30}, \frac{15}{30}, \frac{17}{30}, \frac{19}{30}, \frac{23}{30}, \frac{29}{30}\right\} \neq R_{15}. \tag{28}$$

Acknowledgment. We are grateful to Professor Daubechies, who brought us in touch with each other. Q. Sun also thanks Professor Z. Guo and D. Huang for their encouragement.

REFERENCES

[1] C. K. Chui and J. Z. Wang, *A general framework of compactly supported splines and wavelets*, CAT Report 219, Texas A&M University, College Station, TX, 1990.

[2] A. Cohen, *Ondelettes, analyses multirésolutions et filtres miroirs en quadrature*, Ann. Inst. H. Poincaré, 7 (1990), pp. 439–459.

[3] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[4] W. M. Lawton, *Tight frames of compactly supported affine wavelets*, J. Math. Phys., 31 (1990), pp. 1898–1901.

[5] ———, *Necessary and sufficient conditions for constructing orthonomal wavelet bases*, J. Math. Phys., 32 (1991), pp. 57–61.

[6] S. Mallat, *Multiresolution approximation and wavelet orthonormal bases of $L^2(\mathbb{R})$*, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.

[7] Y. Meyer, *Ondelettes*, Hermann, Paris, 1990.

[8] D. Pollen, *Scaling tiles*, AWARE Inc., 1989.

SIAM J. MATH. ANAL.
Vol. 24, No. 5, pp. 1361–1377, September 1993

015

AN ELEMENTARY APPROACH TO SOME ANALYTIC ASYMPTOTICS*

NICHOLAS PIPPENGER[†]

Abstract. Fredman and Knuth have treated certain recurrences, such as $M(0) = 1$ and

$$M(n+1) = \min_{0 \le k \le n} \big(\alpha M(k) + \beta M(n-k)\big),$$

where $\min(\alpha, \beta) > 1$, by means of auxiliary recurrences such as

$$h(x) = \begin{cases} 0 & \text{if } 0 \le x < 1, \\ 1 + h(x/\alpha) + h(x/\beta) & \text{if } 1 \le x < \infty. \end{cases}$$

The asymptotic behavior of $h(x)$ as $x \to \infty$ with α and β fixed depends on whether $\log\alpha/\log\beta$ is rational or irrational. The solution of Fredman and Knuth used analytic methods in both cases, and used in particular the Wiener–Ikehara Tauberian theorem in the irrational case. The author shows that a more explicit solution to these recurrences can be obtained by entirely elementary methods, based on a geometric interpretation of $h(x)$ as a sum of binomial coefficients over a triangular subregion of Pascal's triangle. Apart from Stirling's formula, in the irrational case only the Kronecker–Weyl theorem (which can itself be proved by elementary methods) is needed, to the effect that if ϑ is irrational, the fractional parts of the sequence $\vartheta, 2\vartheta, 3\vartheta, \ldots$, are uniformly distributed in the unit interval.

Key words. asymptotic analysis, recurrence relation

AMS subject classification. 26A12

1. Introduction.

The analysis of algorithms and data structures, as well as of constructions for systems such as sorting and switching networks, often leads to recurrences. Because recursive algorithms, data structures, and constructions often involve choices that should be made in an optimal way, the recurrences often involve minimization. In their paper, Fredman and Knuth [FK] treat a large number of related recurrences by a combination of combinatorial and analytic methods. The goal of the present paper is to show how in many cases it is possible to replace the analytic component of their solutions with elementary arguments. (Here the terms "analytic" and "elementary" are used in accordance with the practice in number theory: "analytic" refers to methods based on properties of analytic functions of a complex variable, especially residues or integral transforms, while "elementary" refers to the absence of such methods. In particular, "elementary" does not refer to either simplicity or brevity.) As a bonus, we shall see that our analysis leads to a more explicit and informative solution in some cases. A preliminary version of our results appears as [P].

Of the recurrences treated by Fredman and Knuth, the one which best illustrates our contribution is $M(0) = 1$ and

$$M(n+1) = \min_{0 \le k \le n} \big(\alpha M(k) + \beta M(n-k)\big), \tag{1.1}$$

*Received by the editors May 4, 1992; accepted for publication (in revised form) December 3, 1992. This research was partially supported by a Natural Sciences and Engineering Research Council of Canada operating grant and a British Columbia Advance Systems Institute Fellowship Award.

[†] Department of Computer Science, University of British Columbia, Vancouver, British Columbia V6T 1Z2, Canada.

where α and β are fixed parameters with $\min(\alpha, \beta) > 1$. (This is the case "$g(n) = \delta_{n0}$," dealt with in their §6.) By straightforward and elementary arguments, Fredman and Knuth reduce the study of (1.1) to that of the function h defined by

$$h(x) = \begin{cases} 0 & \text{if } 0 \le x < 1, \\ 1 + h(x/\alpha) + h(x/\beta) & \text{if } 1 \le x < \infty. \end{cases} \tag{1.2}$$

The analysis of Fredman and Knuth proceeds by considering the integral transform

$$K(s) = \int_1^\infty \frac{h(t)\,dt}{t^{s+1}}$$

of h, which, with the aid of (1.2), can be shown to be $K(s) = 1/s(1 - \alpha^{-s} - \beta^{-s})$. This function is analytic in the open half plane $\mathrm{Re}(s) > \gamma$, where γ is the unique real solution to

$$\alpha^{-\gamma} + \beta^{-\gamma} = 1. \tag{1.3}$$

Furthermore, $K(s)$ has a simple pole at $s = \gamma$ with residue $C = 1/(\alpha^{-\gamma} \log \alpha^\gamma + \beta^{-\gamma} \log \beta^\gamma)$, as is easily calculated. This pole will ultimately give rise to a factor Cx^γ in the asymptotic behavior of $h(x)$.

The behavior of $K(s)$ on the remainder of the critical line $\mathrm{Re}(s) = \gamma$ depends on whether $\log \alpha / \log \beta$ is rational or irrational. If this quotient is irrational, the pole at $s = \gamma$ is the only one on the critical line, and a Tauberian theorem due to Wiener, Ikehara, and Landau (Lemma 4.3 in Fredman and Knuth) leads to the conclusion that

$$h(x) \sim Cx^\gamma \tag{1.4}$$

in this case. If the quotient $\log \alpha / \log \beta$ is rational, $K(s)$ has additional poles periodically disposed along the critical line. Application of Cauchy's residue theorem leads to the conclusion that

$$h(x) \sim D(x)x^\gamma \tag{1.5}$$

in this case, where $D(x)$ is a periodic function of $\log x$ whose period is determined by the spacing between poles along the critical line, and whose Fourier coefficients are determined by the residues at those poles.

In this paper we shall derive (1.4) and (1.5) in an elementary fashion. This new derivation has the merit of giving a simple explicit formula for the function $D(x)$ in (1.5). We shall also want the solution to the related recurrence

$$h'(x) = \begin{cases} 0 & \text{if } 0 \le x < 1, \\ 1 + \alpha h(x/\alpha) + \beta h(x/\beta) & \text{if } 1 \le x < \infty. \end{cases} \tag{1.6}$$

By analogous elementary methods, we shall show that

$$h'(x) \sim C'x^{\gamma+1}, \tag{1.7}$$

where $C' = 1/(\alpha^{-\gamma} \log \alpha^{\gamma+1} + \beta^{-\gamma} \log \beta^{\gamma+1})$ in the irrational case, and

$$h'(x) \sim D'(x)x^{\gamma+1}, \tag{1.8}$$

where $D'(x)$ is a periodic function of $\log x$ which will be determined explicitly in the rational case.

Fredman and Knuth showed that (1.4) implies that

$$M(n) \sim An^{1+1/\gamma}, \tag{1.9}$$

where A is an explicitly determined constant in the irrational case. We shall show that (1.5) and (1.8) together imply

$$M(n) \sim B(n)n^{1+1/\gamma}, \tag{1.10}$$

where $B(n)$ is an explicitly determined periodic function of $\log n$ in the rational case. (Once the form of the functions $D(x)$ and $D'(x)$ are explicitly known, it is possible to go back and derive these results by extending the analysis of Fredman and Knuth. This would involve showing that certain Fourier series converge to certain periodic functions. But since there is no general procedure for identifying a function from its Fourier series, it does not appear to be possible to extend the analysis of Fredman and Knuth without knowing what $D(x)$ and $D'(x)$ are by some other method.)

2. The rational case. Our analysis begins with the observation that $h(x)$ is the number of words over the alphabet $\{\alpha, \beta\}$ having weight at most x, where the *weight* of a word is the product of its letters. (We take the weight of the empty word to be unity.) Indeed, if $0 \le x < 1$, then there are no such words and $h(x) = 0$. If $1 \le x < \infty$, then $h(x) = 1 + h(x/\alpha) + h(x/\beta)$, and any word for which the product of the letters is at most x must either be empty (and there is 1 such word) or consist of an α followed by a word for which the product of the letters is at most x/α (and there are $h(x/\alpha)$ such words), or consist of an β followed by a word for which the product of the letters is at most x/β (and there are $h(x/\beta)$ such words). Since there are exactly $\binom{i+j}{i}$ words that contain i α's and j β's, we have established the following explicit formula for $h(x)$:

$$h(x) = \sum_{\alpha^i \beta^j \le x} \binom{i+j}{i}. \tag{2.1}$$

Taking logarithms in the constraint of the summation, we see that $h(x)$ may be interpreted as the sum of the binomial coefficients $\binom{i+j}{i}$ in Pascal's triangle over the triangular subregion bounded by the inequalities $i \ge 0$, $j \ge 0$, and

$$i \log \alpha + j \log \beta \le \log x. \tag{2.2}$$

Suppose that $\log \alpha / \log \beta$ is the rational number p/q, where p and q are positive integers such that $\gcd(p, q) = 1$. Then $\log_{\alpha\beta} \alpha = p/(p+q)$, $\log_{\alpha\beta} \beta = q/(p+q)$, and if we set

$$\varrho = (\alpha\beta)^{1/(p+q)},$$

then (2.2) becomes

$$pi + qj \le \log_\varrho x.$$

Since p, q, i, and j are integers, we see that $h(x)$ remains constant as x increases except when $\log_\varrho x$ passes through an integer k when it jumps by

$$S(k) = \sum_{pi+qj=k} \binom{i+j}{i}. \tag{2.3}$$

We shall see below that $S(k)$ has the asymptotic formula

$$S(k) \sim (C \log \sigma)\, \sigma^k, \tag{2.4}$$

where

$$\sigma = \varrho^\gamma = (\alpha\beta)^{\gamma/(p+q)}.$$

If we set $S^*(l) = \sum_{0 \le k \le l} S(k)$, it follows that

$$S^*(l) \sim \frac{(C\sigma \log \sigma)\, \sigma^l}{\sigma - 1}.$$

This formula gives the asymptotic value $S^*(l)$ of $h(x)$ when x is a "magic" number of the form $x = \varrho^l$. The asymptotic formula for arbitrary x follows from this and the fact that $h(x)$ remains constant between magic values of x. If we write $\log_\varrho x = l + \lambda$, where $l = \lfloor \log_\varrho x \rfloor$ (the integral part of $\log_\varrho x$) and $\lambda = \{\log_\varrho x\}$ (the fractional part of $\log_\varrho x$), then

$$\begin{aligned} h(x) &\sim \frac{(C\sigma \log \sigma)\, \sigma^l}{\sigma - 1} \\ &\sim \frac{(C\sigma^{1-\lambda} \log \sigma)\, \sigma^{l+\lambda}}{\sigma - 1} \\ &\sim P(\{\log_\varrho x\})\, x^\gamma, \end{aligned}$$

where

$$P(\lambda) = \frac{C\sigma^{1-\lambda} \log \sigma}{\sigma - 1}.$$

This establishes (1.5) with $D(x) = P(\{\log_\varrho x\})$, which is periodic in $\log x$ (with period $\log \varrho$), as claimed.

It remains to establish (2.4). The major steps of the derivation are as follows. First, we approximate the binomial coefficients $\binom{i+j}{i} = (i+j)!/i!j!$ in (2.3) by applying Stirling's formula to their constituent factorials. If we separate the approximation into algebraically varying factors and exponentially varying factors, we see that the exponentially varying factors impart to the summand a peaking reminiscent of the central limit theorem: the greatest contribution to the sum comes when i and j are in the fixed ratio $\alpha^{-\gamma}/\beta^{-\gamma}$. This variation allows the terms of the sum not near the peak to be neglected. The resulting truncated sum is then estimated by an integral; the error in this estimation is at most the total variation of the summand, which (since the summand is unimodal) is at most twice the largest term. The resulting integral can be transformed into the well-known integral $\int_{-\infty}^{+\infty} e^{-y^2}\, dy = \pi^{1/2}$ by adjoining negligible tails. The result is (2.4).

Successive values of i and j differ by q and p, respectively; it will be convenient to have an index whose successive values differ by 1. Thus we introduce the index m satisfying

$$i = qm, \qquad j = k/q - pm, \qquad i + j = k/q - (p - q)m.$$

This index assumes values that are not necessarily integers, but are congruent to $1/q$ modulo 1.

LEMMA 2.1.

$$\binom{i+j}{i} = \left(1 + O\left(\frac{1}{i} + \frac{1}{j}\right)\right) \left(\frac{i+j}{2\pi i j}\right)^{1/2} \exp kE\left(\frac{m}{k}\right),$$

where

$$E(\mu) = F\big(G(\mu)\big), \qquad F(\nu) = \frac{H(\nu)}{p\nu + q(1-\nu)},$$
$$G(\mu) = \frac{q^2\mu}{1-(p-q)q\mu}, \qquad H(\nu) = -\nu\log\nu - (1-\nu)\log(1-\nu).$$

Proof. The estimate

$$\binom{i+j}{i} = \left(1+O\left(\frac{1}{i}+\frac{1}{j}\right)\right)\left(\frac{i+j}{2\pi ij}\right)^{1/2}\exp(i+j)H\left(\frac{i}{i+j}\right)$$

is an immediate consequence of Stirling's formula

$$n! = \left(1+O\left(\frac{1}{n}\right)\right)(2\pi n)^{1/2}e^{-n}n^n$$

(see Knuth [K1, §1.2.11]). Define ν such that

$$i = \nu(i+j), \qquad j = (1-\nu)(i+j), \qquad k = \big(p\nu + q(1-\nu)\big)(i+j).$$

Then

$$\binom{i+j}{i} = \left(1+O\left(\frac{1}{i}+\frac{1}{j}\right)\right)\left(\frac{i+j}{2\pi ij}\right)^{1/2}\exp kF\left(\frac{i}{i+j}\right).$$

Define μ such that

$$m = \mu k.$$

Then ν and μ are related by

$$\nu = \frac{q^2\mu}{1-(p-q)q\mu}, \qquad \mu = \frac{\nu}{q^2+(p-q)q\nu}.$$

This yields the assertion of the lemma. □

LEMMA 2.2. *The function $F(\nu)$ assumes its unique maximum (for $0 \le \nu \le 1$) at*

$$N = \sigma^{-p}.$$

At this point

$$F(N) = \log\sigma, \qquad F'(N) = 0, \qquad F''(N) = -\frac{1}{N(1-N)\Delta},$$

where

$$\Delta = pN + q(1-N)$$

and the primes indicate differentiation. Accordingly, $E(\mu)$ assumes its maximum at

$$M = \frac{N}{q\Delta},$$

and at this point

$$E(M) = \log\sigma, \qquad E'(M) = 0, \qquad E''(M) = \frac{\Delta^3}{N(1-N)}.$$

Proof. We shall let $H(0) = H(1) = 0$; this makes $H(\nu)$, and therefore $F(\nu)$, continuous on the closed interval $0 \le \nu \le 1$. These functions are in fact analytic in the open interval $0 < \nu < 1$, and thus $F(\nu)$ can assume its maximum only where its first derivative vanishes or at an endpoint. We compute the first derivatives

$$H'(\nu) = \log(1-\nu) - \log\nu,$$
$$F'(\nu) = -\frac{H(\nu)(p-q)}{\left(p\nu + q(1-\nu)\right)^2} + \frac{H'(\nu)}{p\nu + q(1-\nu)}.$$

Equating $F'(\nu)$ to zero leads to the equation

$$\left(p\nu + q(1-\nu)\right)H'(\nu) = (p-q)H(\nu).$$

This has the unique solution $\nu = N$, where

$$N = \alpha^{-\gamma} = \sigma^{-p},$$

so that

$$1 - N = \beta^{-\gamma} = \sigma^{-q}.$$

This gives

$$F(N) = \log\sigma,$$

which is obviously larger than $F(\nu)$ at either of the endpoints. We compute the second derivatives

$$H''(\nu) = -\frac{1}{\nu(1-\nu)},$$
$$F''(\nu) = \frac{2H(\nu)(p-q)^2}{\left(p\nu + q(1-\nu)\right)^3} - \frac{2H'(\nu)(p-q)}{\left(p\nu + q(1-\nu)\right)^2} + \frac{H''(\nu)}{p\nu + q(1-\nu)}.$$

Since the first two terms of $F''(\nu)$ are a multiple of $F'(\nu)$, they vanish at $\nu = N$, leaving

$$F''(N) = -\frac{1}{N(1-N)\Delta}.$$

This derivation can be carried over to $E(\mu)$, $E'(\mu)$, and $E''(\mu)$ through the derivative

$$G'(\mu) = \frac{q^2}{\left(1 - (p-q)q\mu\right)^2} = \left(p\nu + q(1-\nu)\right)^2$$

and the chain rule. □

LEMMA 2.3.

$$\sum_m \left(1 + O\left(\frac{1}{i} + \frac{1}{j}\right)\right)\left(\frac{i+j}{2\pi ij}\right)^{1/2} \exp kE\left(\frac{m}{k}\right)$$
$$= \left(1 + O\left(\frac{(\log k)^{3/2}}{k^{1/2}}\right)\right)\frac{\sigma^k}{\Delta}.$$

Here the sum over m, as described above, is over $i \ge 0$ and $j \ge 0$ such that $i + j = k$, with $m = i/q$.

Proof. The major steps of the derivation are as follows. The central peaking of the summand will be exploited, allowing the tails of the summation to be neglected. The decaudated sum can be simplified, since the algebraically varying factors behave like constants in the remaining range of summation. The resulting sum will be estimated with an integral, to which the tails previously removed will be restored. The recaudated integral can be evaluated by standard methods.

Our sum is

$$\sum_m W_m,$$

where

$$W_m = \left(1 + O\left(\frac{1}{i} + \frac{1}{j}\right)\right)\left(\frac{i+j}{2\pi ij}\right)^{1/2} \exp kE\left(\frac{m}{k}\right).$$

Since $E(\mu)$ is analytic at $\mu = M$, it can be expanded in a Taylor series about this point. The result is

$$E(\mu) = \log\sigma - (\mu - M)^2/\delta^2 + O\big((\mu - M)^3\big),$$

where

$$\delta = \left(\frac{2N(1-N)}{\Delta^3}\right)^{1/2}.$$

We shall break our sum into three parts,

$$\sum_m W_m = \sum_{m<a} W_m + \sum_{a\le m\le b} W_m + \sum_{b<m} W_m,$$

where

$$a = Mk - \left(\frac{6N(1-N)k\log k}{\Delta^3}\right)^{1/2},$$

$$b = Mk + \left(\frac{6N(1-N)k\log k}{\Delta^3}\right)^{1/2}.$$

To estimate the sum over $m < a$, we observe that it comprises $O(k)$ terms, each of which is at most W_a. We have $E(a/k) = \log\sigma - 3(\log k/k) + O\big((\log k/k)^{3/2}\big)$ (by the Taylor expansion). Thus

$$W_a = O\left(\frac{\sigma^k}{k^3}\right),$$

and so

$$\sum_{m<a} W_m = O\left(\frac{\sigma^k}{k^2}\right).$$

Similarly,

$$\sum_{b<m} W_m = O\left(\frac{\sigma^k}{k^2}\right),$$

and thus

$$\sum_m W_m = \sum_{a\le m\le b} W_m + O\left(\frac{\sigma^k}{k^2}\right). \tag{2.5}$$

For any term in the sum over $a \leq m \leq b$,

$$m = Mk\left(1 + O\left(\left(\frac{\log k}{k}\right)^{1/2}\right)\right),$$

from which it follows that

$$i = \frac{Nk}{\Delta}\left(1 + O\left(\left(\frac{\log k}{k}\right)^{1/2}\right)\right),$$

$$j = \frac{(1-N)k}{\Delta}\left(1 + O\left(\left(\frac{\log k}{k}\right)^{1/2}\right)\right),$$

and

$$i + j = \frac{k}{\Delta}\left(1 + O\left(\left(\frac{\log k}{k}\right)^{1/2}\right)\right).$$

Thus

$$W_m = \left(1 + O\left(\frac{(\log k)^{3/2}}{k^{1/2}}\right)\right)\left(\frac{\Delta}{2\pi kN(1-N)}\right)^{1/2} \sigma^k V_m,$$

where

$$V_m = \exp -\left((m - Mk)^2/\delta^2 k\right),$$

and therefore

$$\sum_{a \leq m \leq b} W_m = \left(1 + O\left(\frac{(\log k)^{3/2}}{k^{1/2}}\right)\right)\left(\frac{\Delta^3}{2\pi kN(1-N)}\right)^{1/2} \sigma^k \sum_{a \leq m \leq b} V_m. \tag{2.6}$$

Now,

$$\sum_{a \leq m \leq b} V_m = \int_a^b V_x\,dx + O(1), \tag{2.7}$$

since the total variation of the integrand is $O(1)$. We shall express our integral as the sum of three integrals:

$$\int_a^b V_x\,dx = -\int_{-\infty}^a V_x\,dx + \int_{-\infty}^{+\infty} V_x\,dx - \int_a^{+\infty} V_x\,dx.$$

Integration by parts gives

$$\int_{-\infty}^a V_x\,dx = O\left(\frac{V_a}{a}\right) = O\left(\frac{1}{k^4}\right).$$

Similarly

$$\int_a^{+\infty} V_x\,dx = O\left(\frac{1}{k^4}\right),$$

and thus

$$\int_a^b V_x\,dx = \int_{-\infty}^{+\infty} V_x\,dx + O\left(\frac{1}{k^4}\right). \tag{2.8}$$

Using the transformation

$$x = Mk + \delta k^{1/2} y$$

and the well-known integral

$$\int_{-\infty}^{+\infty} \exp -y^2 \, dy = \pi^{1/2},$$

we obtain

$$\int_{-\infty}^{+\infty} V_x \, dx = \left(\frac{2\pi k N(1-N)}{\Delta^3} \right)^{1/2}.$$

Working backwards through (2.8), (2.7), (2.6), and (2.5) yields the assertion of the lemma. □

The formula (2.4) follows from Lemma 2.3, since $C \log \sigma = 1/\Delta$.

We should mention here that the special cases $p = q = 1$, where $S(k) = 2^k$, and $p = 1$, $q = 2$, where $S(k) = F_{k+1} \sim \phi^{k+1}/\sqrt{5}$ (in which F_n is the nth Fibonacci number and $\phi = (1+\sqrt{5})/2$; see Knuth [K1, §1.2.8, eq. (15) and Exer. 16]) are well known, and the analysis just given can be regarded as a generalization of these cases. Furthermore, that $\sigma = 2$ and $\sigma = \phi$ are algebraic in the examples just cited is not accidental: the rationality of $\log \alpha / \log \beta = p/q$ implies that $\alpha^{-\gamma}$ and $\beta^{-\gamma}$ are roots of the polynomials $(1-z)^p = z^q$ and $z^p = (1-z)^q$, respectively, whence $\sigma = (\alpha\beta)^{\gamma/(p+q)}$ is algebraic.

In §4 we shall also want the solution to the recurrence (1.6) for $h'(x)$ in the rational case. Let us call the product of the letters in a word over the alphabet $\{\alpha, \beta\}$ the *weight* of the word. Then $h'(x)$ is the sum of the weights of all words whose weight is at most x, and thus we have the explicit formula

$$h'(x) = \sum_{\alpha^i \beta^j \le x} \binom{i+j}{i} \alpha^i \beta^j.$$

The treatment of this sum is completely analogous to that of (2.1); the result is

$$h'(x) \sim P'(\{\log_\varrho x\}) \, x^{\gamma+1},$$

where

$$P'(\lambda) = \frac{C' \tau^{1-\lambda} \log \tau}{\tau - 1},$$

in which

$$\tau = \varrho^{\gamma+1} = (\alpha\beta)^{(\gamma+1)/(p+q)}.$$

This establishes (1.8) with $D'(x) = P'(\{\log_\varrho x\})$.

3. The irrational case. When $\log \alpha / \log \beta$ is irrational the analysis of the preceding section is not applicable, for as x increases new binomial coefficients enter the sum one by one, rather than in the regularly spaced platoons of the rational case. Furthermore, the order of their entry is very irregular, with small coefficients near the axes being interspersed with large ones near the main diagonal. The analysis of this section is based on a regularity of averages amid this irregularity of detail, as expressed by the "ergodicity of an irrational rotation of the circle." We shall use in particular the Kronecker–Weyl theorem, to the effect that if ϑ is irrational, then the fractional parts of the sequence $\vartheta, 2\vartheta, 3\vartheta, \ldots$, are uniformly distributed in the unit interval. (This theorem as stated was proved by Weyl [W1]; Kronecker [K2] proved that if ϑ is irrational, then the fractional parts of the sequence $\vartheta, 2\vartheta, 3\vartheta, \ldots$ are dense in the unit interval.) Weyl's orginal proof (which is probably still the simplest proof) of this theorem was based on Fourier series, which by some tastes might not be accepted as elementary. A subsequent proof based on continued fractions (see Nivin [N, Chap. 6, §3]) is incontestably elementary, however.

We shall say that a subset Ξ of the unit interval is an "interval modulo 1" if it is the image modulo 1 of an interval. (Thus $[0, \xi/2) \cup [1-\xi/2, 1)$ is an interval modulo 1 of length ξ.) The Kronecker–Weyl theorem asserts the following.

Let ϑ be irrational. For every $0 < \xi < 1$ and $0 < \eta < 1$, there exists a natural number t such that, if Ξ is an interval modulo 1 of length ξ and T is any set of t consecutive integers, then at least $(1-\eta)\xi t$ and at most $(1+\eta)\xi t$ of the integers i in T are such that $\{i\vartheta\}$ falls in Ξ. (This theorem is often stated in the special case in which $T = \{1, \ldots, t\}$, but shifting T to $T + u$ is equivalent to shifting Ξ to $\Xi - u\vartheta$ modulo 1, so the special case implies the general.)

Let $\varepsilon > 0$ be fixed. Define the function $h_\varepsilon(x)$ by

$$h_\varepsilon(x) = \sum_{xe^{-\varepsilon} < \alpha^i \beta^j \le x} \binom{i+j}{i}. \tag{3.1}$$

Taking logarithms in the constraint of the summation, we see that $h_\varepsilon(x)$ may be interpreted as the sum of the binomial coefficient over the trapezoidal region bounded by the inequalities $i \ge 0$, $j \ge 0$, and

$$\log x - \varepsilon < i \log \alpha + j \log \beta \le \log x. \tag{3.2}$$

We shall see below that $h_\varepsilon(x)$ satisfies the asymptotic inequalities

$$C\gamma\varepsilon(1-\varepsilon)e^{-\gamma\varepsilon}x^\gamma \lesssim h_\varepsilon(x) \lesssim C\gamma\varepsilon(1+\varepsilon)x^\gamma \tag{3.3}$$

as $x \to \infty$ with ε fixed. (Here $f(x) \lesssim g(x)$ means that $\limsup_{x\to\infty} f(x)/g(x) \le 1$.) If we set $l = \lfloor \log x/\varepsilon \rfloor + 1$, then $xe^{-l\varepsilon} < 1$, so we have

$$h(x) = \sum_{0 \le k \le l} h_\varepsilon(xe^{-k\varepsilon}).$$

It follows that

$$\frac{C\gamma\varepsilon(1-\varepsilon)e^{-\gamma\varepsilon}x^\gamma}{1-e^{-\gamma\varepsilon}} \lesssim h(x) \lesssim \frac{C\gamma\varepsilon(1+\varepsilon)x^\gamma}{1-e^{-\gamma\varepsilon}}.$$

Since this holds for every $\varepsilon > 0$, we may let $\varepsilon \to 0$ and obtain (1.4).

It remains to establish (3.3). The proof follows the same general lines as that of (2.4), but is complicated by the fact that the lattice points (i, j) are not equally spaced in the trapezoid (3.2) as they were along the boundary of the triangle (2.2). Our salvation comes from the Kronecker–Weyl theorem, which shows that though they are not "equally spaced," they are "uniformly distributed." This will allow the trapezoid (3.2) to be broken into pieces, each of which is sufficiently large so that it contains a number of lattice points approximately proportional to its area, yet sufficiently small so that the binomial coefficients associated with these lattice points are approximately equal.

Suppose that $\varepsilon < \log\beta$ and set $\vartheta = \log\alpha/\log\beta$, so that ϑ is irrational. Let us say that a natural number i is "lucky" if there exists a j such that i and j satisfy the inequalities (3.2). Clearly there exists such a j if and only if $\{i\vartheta\}$ falls in the interval $\bigl((\log x - \varepsilon)/\log\beta, \log x/\log\beta\bigr]$. (Since the length of this interval is $\varepsilon/\log\beta < 1$, it contains either no integers or one integer.) By the Kronecker–Weyl theorem, with $\xi = \varepsilon/\log\beta$ and $\eta = \varepsilon$, we may choose t such that among any t consecutive natural numbers there are $(1 \pm \varepsilon)\varepsilon t/\log\beta$ lucky values of i.

For lucky i, we shall regard j as a function of i. We shall abbreviate $i\log\alpha + j\log\beta$ by k (which is not necessarily an integer). We shall abbreviate $\log x$ by l (which is not necessarily an integer), and $l - k$ by λ (so that $0 \le \lambda < \varepsilon$).

Since we no longer have the parameters p and q, we shall use $\log\alpha$ and $\log\beta$ in their stead. Thus we introduce m satisfying

$$i = m\log\beta, \quad j = (l-\lambda)/\log\beta - m\log\alpha, \quad i + j = (l-\lambda)/\log\beta - m(\log\alpha - \log\beta).$$

Let us say that a value of m is "lucky" if it corresponds to a lucky value of i. Henceforth, we shall take m to range over lucky values, and regard i, j, k, and λ as functions of m for these lucky values.

By analogy with Lemma 2.1, we have

$$\binom{i+j}{i} = \left(1 + O\left(\frac{1}{i} + \frac{1}{j}\right)\right)\left(\frac{i+j}{2\pi ij}\right)^{1/2} \exp kE\left(\frac{m}{k}\right),$$

where

$$E(\mu) = F\bigl(G(\mu)\bigr), \qquad F(\nu) = \frac{H(\nu)}{\nu\log\alpha + (1-\nu)\log\beta},$$
$$G(\mu) = \frac{\mu(\log\beta)^2}{1 - (\log\alpha - \log\beta)\log\beta}, \qquad H(\nu) = -\nu\log\nu - (1-\nu)\log(1-\nu).$$

By analogy with Lemma 2.2, the function $F(\nu)$ assumes its unique maximum (for $0 \le \nu \le 1$) at

$$N = \alpha^{-\gamma}, \qquad 1 - N = \beta^{-\gamma}.$$

At this point

$$F(N) = \gamma, \qquad F'(N) = 0, \qquad F''(N) = -\frac{1}{N(1-N)\Delta},$$

where

$$\Delta = N\log\alpha + (1-N)\log\beta.$$

Accordingly, $E(\mu)$ assumes its maximum at

$$M = \frac{N}{\Delta\log\beta},$$

and at this point

$$E(M) = \gamma, \qquad E'(M) = 0, \qquad E''(M) = \frac{\Delta^3}{N(1-N)}.$$

We now seek the analog of Lemma 2.3, which is the following.

LEMMA 3.1.

$$\begin{aligned}&\left(1+O\left(\frac{(\log l)^{3/2}}{l^{1/2}}\right)\right)\frac{(1-\varepsilon)e^{-\gamma\varepsilon}\varepsilon x^{\gamma}}{\Delta}\\ &\leq\sum_m\left(1+O\left(\frac{1}{i}+\frac{1}{j}\right)\right)\left(\frac{i+j}{2\pi ij}\right)^{1/2}\exp kE\left(\frac{m}{k}\right)\\ &\leq\left(1+O\left(\frac{(\log l)^{3/2}}{l^{1/2}}\right)\right)\frac{(1+\varepsilon)\varepsilon x^{\gamma}}{\Delta}.\end{aligned}$$

In this lemma and all that follows, the constants .implicit in the O-terms may depend on ε. The inequalities involving the O-terms are to be interpreted as follows: for every $\varepsilon > 0$ and every choice of the O-term in the middle expression, there exist choices of the O-terms in the outer expression such that the inequalities are satisfied for all x.

Proof. Since $l-\varepsilon \leq k \leq l$ and $E(\mu) \leq \gamma$, we have

$$e^{-\gamma\varepsilon}\exp lE\left(\frac{m}{k}\right) \leq \exp kE\left(\frac{m}{k}\right) \leq \exp lE\left(\frac{m}{k}\right).$$

Thus it will suffice to prove

$$\begin{aligned}&\left(1+O\left(\frac{(\log l)^{3/2}}{l^{1/2}}\right)\right)\frac{(1-\varepsilon)\varepsilon x^{\gamma}}{\Delta}\\ &\leq\sum_m\left(1+O\left(\frac{1}{i}+\frac{1}{j}\right)\right)\left(\frac{i+j}{2\pi ij}\right)^{1/2}\exp lE\left(\frac{m}{k}\right)\\ &\leq\left(1+O\left(\frac{(\log l)^{3/2}}{l^{1/2}}\right)\right)\frac{(1+\varepsilon)\varepsilon x^{\gamma}}{\Delta},\end{aligned}\tag{3.4}$$

Choose t using the Kronecker–Weyl theorem so that, for any interval of length $\varepsilon/\log\beta$ modulo 1, among any t consecutive integers i, there are between $(1-\varepsilon)\varepsilon t/\log\beta$ and $(1+\varepsilon)\varepsilon t/\log\beta$ such that $\{i\vartheta\}$ falls in the given interval modulo 1. Set $L = (t\log\beta)/2$. Set

$$Q=\left(\frac{6N(1-N)l\log l}{\Delta^3}\right)^{1/2},$$

and set

$$q=\left\lceil\frac{Q-L}{2L}\right\rceil,$$

so that $(2q+1)L$ is the smallest odd multiple of L that is not less than Q.

Our sum is

$$\sum_m W_m,$$

where

$$W_m=\left(1+O\left(\frac{1}{i}+\frac{1}{j}\right)\right)\left(\frac{i+j}{2\pi ij}\right)^{1/2}\exp lE\left(\frac{m}{k}\right).$$

Since $E(\mu)$ is analytic at $\mu = M$, it can be expanded in a Taylor series about this point. The result is

$$E(\mu)=\gamma-(\mu-M)^2/\delta^2+O\big((\mu-M)^3\big),$$

where

$$\delta = \left(\frac{2N(1-N)}{\Delta^3}\right)^{1/2}.$$

We shall break our sum into three parts,

$$\sum_m W_m = \sum_{m<a} W_m + \sum_{a\le m<b} W_m + \sum_{b\le m} W_m,$$

where

$$a = Ml - (2q+1)L,$$
$$b = Ml + (2q+1)L.$$

We shall need the following approximation property of $E(\mu)$. If m is of the form

$$m = Ml + O\big((l\log l)^{1/2}\big),$$

then (since $|k-l| \le \varepsilon$) we have

$$\left|\frac{m}{k} - \frac{m}{l}\right| = O\left(\frac{1}{l}\right).$$

Furthermore, since $E'(\mu) = O\big(|\mu - M|\big)$ in a fixed neighborhood of M, we have

$$\left|E\left(\frac{m}{k}\right) - E\left(\frac{m}{l}\right)\right| = O\left(\frac{(\log l)^{1/2}}{l^{3/2}}\right).$$

To estimate the sum over $m < a$, we observe that it comprises $O(l)$ terms, each of which is at most W_a. We have $E(a/k) = E(a/l) + O\big((\log l)^{1/2}/l^{3/2}\big)$ (by the approximation property with $m = a$) and $E(a/l) = \gamma - 3(\log l/l) + O\big((\log l/l)^{3/2}\big)$ (by the Taylor series expansion). Thus

$$W_a = O\left(\frac{x^\gamma}{l^3}\right),$$

and so

$$\sum_{m<a} W_m = O\left(\frac{x^\gamma}{l^2}\right).$$

Similarly,

$$\sum_{b\le m} W_m = O\left(\frac{x^\gamma}{l^2}\right),$$

and thus

$$\sum_m W_m = \sum_{a\le m<b} W_m + O\left(\frac{x^\gamma}{l^2}\right). \tag{3.5}$$

For any term in the sum over $a \le m < b$,

$$m = Ml\left(1 + O\left(\left(\frac{\log l}{l}\right)^{1/2}\right)\right),$$

from which it follows that

$$i = \frac{Nl}{\Delta}\left(1 + O\left(\left(\frac{\log l}{l}\right)^{1/2}\right)\right),$$

$$j = \frac{(1-N)l}{\Delta}\left(1 + O\left(\left(\frac{\log l}{l}\right)^{1/2}\right)\right),$$

and

$$i + j = \frac{l}{\Delta}\left(1 + O\left(\left(\frac{\log l}{l}\right)^{1/2}\right)\right).$$

Thus

$$W_m = \left(1 + O\left(\frac{(\log l)^{3/2}}{l^{1/2}}\right)\right)\left(\frac{\Delta}{2\pi l N(1-N)}\right)^{1/2} x^\gamma V_m,$$

where

$$V_m = \exp -\left((m - Mk)^2/\delta^2 k\right).$$

And, therefore,

$$\sum_{a \le m < b} W_m = \left(1 + O\left(\frac{(\log l)^{3/2}}{l^{1/2}}\right)\right)\left(\frac{\Delta^3}{2\pi l N(1-N)}\right)^{1/2} x^\gamma \sum_{a \le m < b} V_m. \tag{3.6}$$

To estimate this sum, we divide the interval $[a, b)$ into $2q + 1$ intervals, each of length $2L$:

$$\sum_{a \le m < b} V_m = \sum_{-q \le r \le q} \sum_{m \in I_r} V_m,$$

where $I_r = \left[Ml + (2r-1)L, Ml + (2r+1)L\right)$ is the half-open interval of length $2L$ centered at $c_r = Ml + 2rL$.

We shall need the following approximation property of V_m. If m and c are each of the form

$$m, c = Ml + O\left((l \log l)^{1/2}\right),$$

and $|m - c| \le L$, then we have

$$\left|\frac{m}{l} - \frac{c}{l}\right| = O\left(\frac{1}{l}\right).$$

Furthermore, since $(d/dm)\log V_m = O\left(|\mu - M|\right)$ in a fixed neighborhood of M, we have

$$V_m = V_c\left(1 + O\left(\frac{(\log l)^{1/2}}{l^{3/2}}\right)\right).$$

Using this approximation property, we may replace the summand V_m by the constant V_{c_r} in the inner sum over $m \in I_r$, so that

$$\sum_{m \in I_r} V_m = \left(1 + O\left(\frac{(\log l)^{1/2}}{l^{3/2}}\right)\right) V_{c_r} \sum_{m \in I_r} 1.$$

By the Kronecker–Weyl theorem, we have

$$(1-\varepsilon)\varepsilon 2L \le \sum_{m\in I_r} 1 \le (1+\varepsilon)\varepsilon 2L$$

for each r, since the lucky values of m in I_r correspond to lucky values of i in an interval of length t. Thus we have

$$(1-\varepsilon)\varepsilon 2L \sum_{-q\le r\le q} V_{c_r} \le \sum_{a\le m<b} V_m \le (1+\varepsilon)\varepsilon 2L \sum_{-q\le r\le q} V_{c_r}. \tag{3.7}$$

The sum $\sum_{-q\le r\le q} V_{c_r}$ may now be approximated by an integral, extended to an infinite range of integration, and evaluated by an appropriate substitution, all as in the proof of Lemma 2.3. The result is

$$\sum_{-q\le r\le q} V_{c_r} = \frac{1}{2L}\left(\frac{2\pi l N(1-N)}{\Delta^3}\right)^{1/2} + O\left(\frac{1}{l^4}\right).$$

Working backwards through (3.7), (3.6), and (3.5) yields (3.4). □

The formula (3.3) follows from Lemma 3.1, since $C\gamma = 1/\Delta$. We observe that the same method works to establish the asymptotic formula (1.7) for $h'(x)$ in the irrational case.

Though we have derived (1.4) and (1.5) by parallel arguments, there is an important difference between these derivations. We could have done the analysis in §2 to obtain an O-estimate for the error in (1.5); the most straightforward way of doing this yields a factor of $\big(1 + O\big((\log\log x)^{3/2}(\log x)^{-1/2}\big)\big)$. No such sharpening is possible for (1.4), however, since the Kronecker–Weyl theorem, in the form we have cited, gives no estimate for the rate of convergence to the uniform distribution. The same phenomenon arises for the analytic proof using the Wiener–Ikehara theorem, for while convergence follows from the behavior of $K(s)$ on the critical line and the right half plane it bounds, the rate of convergence depends on how closely the poles in the left half plane approach the critical line as their imaginary parts grow (see [W2], [I], [L]). With either method, the missing information depends on how well the irrational number $\log\alpha/\log\beta$ can be approximated by rational numbers as the denominators of these rational numbers grow. This is the crux of the difference: all rational numbers are alike, but each irrational number is irrational in its own way.

Since we have not made any quantitative hypothesis concerning the irrationality of $\log\alpha/\log\beta$, we cannot expect to draw any conclusion about the rate of approach in (1.4). If however we assume that $|\log\alpha/\log\beta - p/q|$ is bounded away from zero by a function of q, the elementary method used here (as well as the analytic method used by Fredman and Knuth) can be adapted to yield an explicit O-estimate in (1.4).

4. Conclusion. After deriving (1.4) and (1.5) in a new way, and obtaining explicit descriptions of the functions $D(x)$ and $D'(x)$ appearing in (1.5) and (1.8), we shall exhibit in this section the consequences of these explicit descriptions for the original recurrence (1.1).

Fredman and Knuth show, by elementary arguments, that

$$M(n) = 1 + (\alpha+\beta-1)W(n), \tag{4.1}$$

where $W(n)$ is the sum of the weights of the n words having the smallest weights. (Recall that the weight of a word over $\{\alpha,\beta\}$ is the product of its letters.) By the definitions of $h(x)$ and $h'(x)$, we have $W\big(h(x)\big) = h'(x)$. Let us assume that $\log\alpha/\log\beta$

is rational. Recall that a value of x is "magic" if $x = \varrho^l$ for some natural number l. We have $D(x) = P(0)$ and $P'(x) = P'(0)$ for all magic values of x, and the asymptotic formulas

$$h(x) \sim P(0)\, x^\gamma \tag{4.2}$$

and

$$h'(x) \sim P'(0)\, x^{\gamma+1}, \tag{4.3}$$

valid for magic values of x.

Let us say that a value of n is "magic" if $n = h(x)$ for some magic value of x. Then (4.2) and (4.3) yields the asymptotic formula

$$W(n) \sim P'(0) \left(\frac{n}{P(0)}\right)^{1+1/\gamma}, \tag{4.4}$$

valid for magic values of n.

To extend (4.4) to arbitrary values of n, we observe that as n increases between magic values, $W(n)$ increases by the addition of equal weights. Thus the points of the graph of $W(n)$ between magic values of n lie on the chords joining the points of the graph at successive magic values, and the formula for arbitrary n is obtained by linearly interpolating between the values given by (4.4) for magic values of n. This gives

$$W(n) \sim P'(0) Q\left(\left\{\log_\sigma \left(\frac{n}{P(0)}\right)\right\}\right) \left(\frac{n}{P(0)}\right)^{1+1/\gamma},$$

where $Q(\lambda) = (1 - \lambda + \lambda\tau)\tau^{-\lambda}$, which establishes (1.10) with

$$B(n) = (\alpha + \beta - 1) P'(0) Q\left(\left\{\log_\sigma \left(\frac{n}{P(0)}\right)\right\}\right) \left(\frac{1}{P(0)}\right)^{1+1/\gamma},$$

which is periodic in $\log n$ (with period $\log \sigma$), as claimed.

We have dealt in this paper with particular recurrences, (1.1) and (1.2), taken from Fredman and Knuth. It is possible to extend the analysis straightforwardly to a number of other recurrences of similar form, as, for example, with initial conditions imposed on an initial segment $\{0, 1, \ldots, r\}$ of the domain, rather than just at the point zero, or with three terms on the right-hand side, rather than just two. We see the contribution of this paper, however, as residing more in its methods than in their scope. The Wiener–Ikehara–Landau theorem used by Fredman and Knuth is of an essentially "Tauberian" character, inferring the asymptotics of a sequence from that of its sum. It is virtually equivalent in depth to the prime number theorem, which was in fact the application that motivated Wiener, Ikehara, and Landau. The arguments used in this paper, however, are not only elementary, but also "direct" or "Abelian" in character: they infer the asymptotics of a sum from that of its terms. These arguments are much less delicate than the ones they replace, and they show the phenomena we have studied to be less deep than has hitherto been supposed.

REFERENCES

[FK] M. L. Fredman and D. E. Knuth, *Recurrence relations based on minimization*, J. Math. Anal. Appl., 48 (1974), pp. 534–559.

[I] S. IKEHARA, *An extension of Landau's theorem in the analytical theory of numbers*, J. Math. Phys., 10 (1931), pp. 1–12.

[K1] D. E. KNUTH, *The Art of Computer Programming–Volume* 1: *Fundamental Algorithms*, Addison-Wesley, Reading, MA, 1968.

[K2] L. KRONECKER, *Näherungsweisse Ganzzahlige Auflösung Linearer Gleichungen*, Sitzungsber. Berliner-Akad. Wiss., 46 (1884), pp. 1071 ff.

[L] E. LANDAU, *Über den Wienerschen neuen Weg zum Primzahlsatz*, Sitzber. Preussische Akad. Wiss., (1932), pp. 514–521.

[N] I. NIVEN, *Irrational Numbers*, Mathematical Association of America, Washington, D.C., 1956.

[P] N. PIPPENGER, *An Elementary Approach to Some Analytic Asymptotics*, Proc. Scand. Workshop on Algorithm Theory, 3 (1992), pp. 53–61.

[W1] H. WEYL, *Über die Gleichverteilung von Zahlen modulo Eins*, Math. Annalen, 77 (1916), pp. 313–352.

[W2] N. WIENER, *A new method in Tauberian theorems*, J. Math. Phys., 7 (1928), pp. 161–184.

SIAM J. MATH. ANAL.
Vol. 24, No. 5, pp. 1378–1393, September 1993

TIME-FREQUENCY LOCALIZATION VIA THE WEYL CORRESPONDENCE*

JAYAKUMAR RAMANATHAN† AND PANKAJ TOPIWALA‡

Abstract. A technique of producing signals whose energy is concentrated in a given region of the time-frequency plane is examined. The degree to which a particular signal is concentrated is measured by integrating the Wigner distribution over the given region. This procedure was put forward by Flandrin, and has been used for time-varying filtering in the recent work of Hlawatsch, Kozek, and Krattenthaler. In this paper, the associated operator is studied. Estimates for the eigenvalue decay and the smoothness and decay of the eigenfunctions are established.

Key words. time frequency localization, Weyl correspondence, Wigner distribution

AMS subject classifications. 45A05, 45C05, 42A60, 94A11

1. Introduction.

It is well known that the time-frequency characteristics of a square integrable signal cannot be arbitrary. For example, no such signal can be both time and band limited. The Heisenberg uncertainty principle provides another quantitive restriction on the joint time-frequency behavior of a square integrable signal. These facts indicate that a signal cannot have all its energy concentrated in a finite region of the time-frequency plane.

Nevertheless, in many applications it is important to use signals whose time-frequency characteristics are highly localized. Among other things, the work of Landau, Pollack, and Slepian [7], [8] produced a rigorous development of band-limited signals that are as concentrated as possible within a prescribed timespan. More recently, there has been interest in finding signals that are localized in general regions of phase space via methods that keep time and frequency on an equal footing (see the papers of Daubechies and Paul [5], [6]).

In this paper, we study a localization technique that uses the Wigner distribution to measure the degree to which a signal is concentrated in a particular region. This leads to a self-adjoint *localization* operator that is easy to study in terms of the Weyl correspondence. Under the Weyl correspondence, a function of two variables—called the *symbol*—is associated with an operator on functions of one variable. The symbol of the localization operator is simply the indicator function of the given region in the time-frequency space. The eigenfunctions of this operator with large eigenvalues span a subspace that can be used to determine the component of a general signal that is concentrated within the given region of the time-frequency plane: one computes the projection of a general signal into this subspace. This procedure was put forward by Flandrin [10], who derived a number of useful results, including Lemma 4 below. It has since been developed in the context of time-varying filtering (see, for example, the

*Received by the editors August 22, 1991; accepted for publication (in revised form) November 12, 1992.

†Department of Mathematics, Eastern Michigan University, Ypsilanti, Michigan 48197. This author was partially supported by the MITRE Sponsored Research program and by a Faculty Research and Creative Projects Fellowship Award from Eastern Michigan University during various stages of this work.

‡MITRE Corporation, Mail Stop E025, 202 Burlington Road, Bedford, Massachusetts 01730. This author was supported by the MITRE Sponsored Research program.

papers of Hlawatsch, Kozek, and Krattenthaler [2], [3]). This paper is devoted to the further study of the localization operator descibed above. The asymptotic properties of the eigenvalues are studied and the fact that they are $\mathcal{O}(k^{-3/4})$ is established. In addition, the eigenfunctions with nonzero eigenvalues are shown to have faster than exponential decay in both the time and frequency domains. This, of course, leads to a statement regarding the smoothness of these eigenfunctions. In particular, we show that they are analytic. In the last section, some numerical examples are provided.

2. The Weyl correspondence.

The basic properties of the Weyl correspondence that we will need are collected in this section; for a thorough treatment, the reader is referred to the book of Folland [1].

Let $f, g \in L^2(\mathbb{R})$. A general time-frequency shift of f is

$$\rho(\tau,\sigma)f(t) = e^{\pi i\tau\sigma}e^{2\pi i\sigma t}f(t+\tau).$$

The cross-ambiguity function of f and g is

$$\begin{aligned} A(f,g)(\tau,\sigma) &= \langle \rho(\tau,\sigma)f, g\rangle \\ &= \int e^{\pi i\tau\sigma}e^{2\pi i\sigma t}f(t+\tau)\overline{g(t)}\,dt \\ &= \int e^{2\pi i\sigma s}f(s+\tau/2)\overline{g(s-\tau/2)}\,ds. \end{aligned}$$

The value of $A(f,g)$ at a particular point is the cross-correlation between a particular time-frequency shift of f against g. The ambiguity function is therefore a time-frequency cross-correlation between two functions. The Wigner distribution is the two-dimensional Fourier transform of the cross ambiguity function, thus giving it the interpretation of a time-varying spectrum. The Wigner distribution can be written as

$$W(f,g)(\xi,t) = \int e^{-2\pi i\tau\xi}f(t+\tau/2)\overline{g(t-\tau/2)}\,d\tau.$$

Several useful properties of the Wigner distribution are catalogued below.

THEOREM 1. *Let $f, g \in L^2(\mathbb{R})$. Then*

(1) $W(f,g)(\xi,t) \in L^2(\mathbb{R}^2)$ *and* $\|W(f,g)\|^2 = \|f\|^2\|g\|^2$,
(2) $W(f,g) \in C_0(\mathbb{R}^2)$ *and* $\|W(f,g)\|_\infty \le \|f\|\|g\|$,
(3) $W(g,f) = \overline{W(f,g)}$,
(4) $W(\hat{f},\hat{g})(\xi,t) = W(f,g)(t,-\xi)$, *and*
(5) $W(\rho(a,b)f, \rho(a,b)g)(\xi,t) = W(f,g)(\xi-b, t+a)$.

The Weyl correspondence uses the Wigner distribution to define a correspondence between functions of two variables and operators on $L^2(\mathbb{R})$. It is defined, via duality, by

$$\langle L_S f, g\rangle = \iint S(\xi,t)W(f,g)(\xi,t)\,d\xi dt,$$

where $f, g \in L^2(\mathbb{R})$ and $S(\xi,t)$ is a function with appropriate decay properties. $S(\xi,t)$ is the symbol of the operator L_S. The following theorem of Pool [9] is useful.

THEOREM 2. *A symbol $S(\xi,t) \in L^2(\mathbb{R}^2)$ gives rise to an operator L_S that is Hilbert–Schmidt on $L^2(\mathbb{R})$. Moreover, the mapping $S \mapsto L_S$ is a unitary operator from $L^2(\mathbb{R}^2)$ to the Hilbert–Schmidt operators on $L^2(\mathbb{R})$.*

LEMMA 3. *L_S is self-adjoint if $S(\xi, t)$ is real valued.*

The following result was derived by Flandrin in [10], based on results of Janssen [11]. We will follow the development in [1].

LEMMA 4. *The eigenfunctions of the operator L_S corresponding to a radially symmetric symbol S are:*

$$h_j(t) = \frac{2^{1/4}}{\sqrt{j!}} \left(\frac{-1}{2\sqrt{\pi}}\right)^j e^{\pi t^2} \frac{d^j}{dt^j} e^{-2\pi t^2}.$$

Proof. Theorem 1.105 in [1] shows that

$$W(h_j, h_k)(\xi, t) = \begin{cases} 2(-1)^k \sqrt{\frac{k!}{j!}} e^{-2\pi|z|^2} (2\sqrt{\pi} z)^{j-k} L_k^{(j-k)}(4\pi|z|^2) & \text{for } j \geq k, \\ 2(-1)^j \sqrt{\frac{j!}{k!}} e^{-2\pi|z|^2} (2\sqrt{\pi}\bar{z})^{k-j} L_j^{(k-j)}(4\pi|z|^2) & \text{for } k \geq j, \end{cases}$$

where $z = t + i\xi$ and $L_k^{(\alpha)}$ is the associated Laguerre polynomial. Set $r = |z|$ and note that

$$\begin{aligned} \langle L_S h_j, h_k \rangle &= \iint S(r) W(h_j, h_k)\, d\xi dt \\ &= \begin{cases} 0 & \text{for } j \neq k, \\ (-1)^j 4\pi \int_0^\infty S(r)\, e^{-2\pi r^2} L_j^{(0)}(4\pi r^2)\, r dr & \text{for } j = k. \end{cases} \end{aligned}$$

Hence

$$\lambda_j = (-1)^j 4\pi \int_0^\infty S(r)\, e^{-2\pi r^2} L_j^{(0)}(4\pi r^2)\, r dr. \tag{1}$$

3. Localization via a cut-off. The localization operator we are concerned with is L_{χ_Ω}, where χ_Ω is the characteristic function of some bounded domain in the $t - \xi$ plane. We will assume that $\Omega \subset [-B, B] \times [-T, T]$. Note that

$$\langle L_{\chi_\Omega} f, g \rangle = \iint_\Omega W(f, g).$$

L_{χ_Ω} is self-adjoint since χ_Ω is real valued. Pool's theorem implies that L_{χ_Ω} is Hilbert–Schmidt. Hence there is an orthonormal basis $\phi_1, \phi_2, \ldots$ of $L^2(\mathbb{R})$ and real numbers $\lambda_1, \lambda_2, \ldots$ such that $L_{\chi_\Omega}\phi_k = \lambda_k \phi_k$. The Hilbert–Schmidt norm of L_{χ_Ω} is $\sum |\lambda_k|^2 = |\Omega|$, the measure of Ω. We will assume that the eigenvalues are arranged in order of decreasing absolute value. It is easy to check that the largest positive eigenvalue corresponds to the maximum energy an L^2 function can have within the domain Ω. The corresponding eigenfunction would then be a time function with energy as concentrated as possible within Ω. Our principal aim in this paper is to study the decay of the eigenvalues of L_{χ_Ω} and the smoothness properties of the eigenfunctions.

Several properties of the associated kernel will be of importance.

LEMMA 5. *The kernel of the operator L_{χ_Ω} is given by the equation*

$$K(s,t) = \int \chi_\Omega\left(\xi, \frac{s+t}{2}\right) e^{2\pi i(s-t)\xi}\, d\xi$$

and has the properties

(1) $K(s,t) \equiv 0$ *if* $|s+t| \geq 2T$; *and*

(2) *if the cross-sections* Ω_t *of* Ω *in the* ξ *direction consist of at most* M *intervals, then* $|K(s,t)| \leq CM/(|s|+1)$.

Proof. The formula for the kernel is well known (see [1]). The kernel can be written as

$$K(s,t) = F\left(t-s, \frac{s+t}{2}\right),$$

where

$$F(\eta,t) = \int_{\Omega_t} e^{-2\pi i\eta\xi}\, d\xi.$$

Item (1) follows from the observation that $\Omega_t = \emptyset$ if $|t| \geq T$. We now verify item (2). By assumption, the cross-section is the union of at most M disjoint intervals: $\Omega_t = \cup_{i=1}^{L}[\alpha_i, \beta_i]$ for some integer $L \leq M$. Therefore, one can estimate that

$$\begin{aligned}|F(\eta,t)| &= \left|\sum e^{-\pi i(\alpha_k+\beta_k)\eta} \frac{\sin(\pi(\beta_k-\alpha_k)\eta)}{\pi\eta}\right| \\ &\leq \sum \left|\frac{\sin(\pi(\beta_k-\alpha_k)\eta)}{\pi\eta}\right| \\ &\leq \frac{2L}{|\eta|+1} \leq \frac{2M}{|\eta|+1}\end{aligned}$$

since $\sin(Ax)/x \leq \max(2,A)/(|x|+1)$. The estimate only needs to be verified when $|s+t| \leq 2T$. In this case,

$$|s-t| + 2T \geq |s-t| + |s+t| \geq 2|s|.$$

Using this and the estimate for $F(\eta,t)$ yields

$$\begin{aligned}|K(s,t)| &\leq \frac{CM}{|s-t|+1} \\ &\leq \frac{CM}{2|s|+1-2T}\end{aligned}$$

for all large s. The estimate in the theorem follows easily by adjusting the constant as necessary.

For domains Ω with piecewise C^1 boundary, we can show that λ_k is $\mathcal{O}(k^{-3/4})$. The proof is a modification of Weyl's classical work on the asymptotics of eigenvalues of integral equations. The following lemma contains some useful standard facts about the Weyl correspondence. Again we refer the reader to [1] for the proofs.

LEMMA 6. (1) *The operators corresponding to $S(\xi,t)$ and $S(\xi-\xi_0,t-t_0)$ are unitarily equivalent.*

(2) *Suppose the symbols $S_1(\xi,t)$ and $S_2(\xi,t)$ are related by an orthogonal change of variables. Then the corresponding operators are unitarily equivalent.*

We will first prove that λ_k is $\mathcal{O}(k^{-3/4})$ for domains Ω of the following form:

$$(*) \qquad \Omega = \{(\xi,t) : \alpha(t) \le \xi \le \beta(t), t \in [-T,T]\},$$

where $\alpha(t),\beta(t)$ are C^1 functions with vanishing endpoint values. The kernel associated to the symbol $\chi_\Omega(\xi,t)$ is

$$K(s,t) = e^{-\pi i(t-s)[\alpha+\beta](\frac{s+t}{2})}\,\frac{\sin(\pi(t-s)[\beta-\alpha](\frac{s+t}{2}))}{\pi(t-s)}.$$

Moreover, it is easy to check that K is a Lipschitz function whose gradient DK exists almost everywhere (a.e.) and satisfies the inequality

$$(2) \qquad |DK| \le C \quad \text{a.e.}$$

LEMMA 7. *If Ω is of the form $(*)$, then there are symmetric finite rank approximations*

$$k_N(s,t) = \sum a_Q\chi_Q,$$

where Q ranges over squares of the form

$$\{(\xi,t) : i/N \le \xi \le (i+1)/N, j/N \le t \le (j+1)/N\}, \qquad -N^2 \le i,j \le N^2-1$$

with the property that

$$(3) \qquad \iint |K(s,t)-k_N(s,t)|^2\,dsdt \quad \textit{is } \mathcal{O}\left(\frac{1}{N}\right).$$

Proof. Set

$$a_Q = \frac{1}{|Q|}\iint_Q K(s,t)\,dsdt.$$

The symmetry of the kernel K forces symmetry of k_N. Let $R = \cup Q$. Then

$$\begin{aligned}\iint_R |K(s,t)-k_N(s,t)|^2\,dsdt &= \sum\iint_Q |K(s,t)-k_N(s,t)|^2\,dsdt \\ &\le \sum |Q|\iint_Q |DK|^2 = \frac{1}{N^2}\iint_R |DK|^2\end{aligned}$$

using Poincaré's inequality [4]. Since $K(s,t)$ is supported within the strip $|s+t| \le 2T$, (2) yields the estimate

$$(4) \qquad \iint_R |K(s,t)-k_N(s,t)|^2\,dsdt \le \frac{C}{N}.$$

The mean square error over the exterior of R can be handled as follows:

$$(5) \qquad \iint_{\mathbb{R}^2-R} |K-k_N|^2 = \iint_{\mathbb{R}^2-R} |K|^2\,dsdt \le \frac{C}{N},$$

in view of Lemma 5. Putting (4) and (5) together yields the estimate in (3).

In view of Satz III of Weyl's paper [4], we have

$$\lambda^2_{4N^2+1} + \cdots \leq \frac{C}{N}.$$

We will apply this inequality to $N/2$, where $N = [\sqrt{k-1}/2]$. (Here [ ] denotes the greatest integer function.) This yields

$$\lambda^2_{4(N/2)^2-1} + \cdots + \lambda^2_{4N^2-1} + \cdots + \lambda^2_k \cdots \leq \frac{2C}{N}.$$

Since the eigenvalues have been arranged in decreasing order, the left-hand side of the above inequality is greater than or equal to $3N^2\lambda^2_k$. This yields an inequality of the form $\lambda^2_k \leq C/N^3$. Clearly $1/N$ is $\mathcal{O}(1/\sqrt{k})$. These remarks imply that λ_k is $\mathcal{O}(k^{-3/4})$ for domains with property (*). We now use a cut and paste argument to derive the result for general Ω with C^1 boundary.

LEMMA 8. *If $K'(s,t)$ and $K''(s,t)$ are two symmetric real kernels in $L^2(\mathbb{R}^2)$ with*

$$|\lambda'_k| \leq \frac{C'}{k^{3/4}} \quad \text{and} \quad |\lambda''_k| \leq \frac{C''}{k^{3/4}},$$

then the eigenvalues of $K(s,t) = K'(s,t) + K''(s,t)$ must satisfy the same estimate: $|\lambda_k| \leq Ck^{-3/4}$ for some constant C.

Proof. Satz I of Weyl's paper [4] implies that

$$|\lambda_{2k}| \leq |\lambda'_k| + |\lambda''_{k+1}|,$$
$$|\lambda_{2k+1}| \leq |\lambda'_{k+1}| + |\lambda''_{k+1}|$$

for all positive integers k. From this it is straightforward to verify that the eigenvalues of $K(s,t)$ have the required decay property.

THEOREM 9. *If Ω is a bounded domain with piecewise C^1 boundary, then λ_k is $\mathcal{O}(k^{-3/4})$.*

Proof. Clearly, such an Ω can be decomposed into a finite union of nonoverlapping subdomains $\Omega = \cup_k \Omega_k$, where each Ω_k can be put into the form (*) after a rigid motion in the plane. By Lemma 6, each $L_{\chi_{\Omega_k}}$ has eigenvalues with the sought-after decay property. Lemma 8 then implies that L_{χ_Ω} also has the same property.

This estimate is in fact sharp, at least for annular regions.

PROPOSITION 10. *Let $\Omega = \{(\xi,t) : \epsilon \leq (\xi^2+t^2)^{1/2} \leq R\}$, where $0 < \epsilon < R$. Then*

$$0 < \limsup_{k\to\infty} k^{3/4}\lambda_k < \infty.$$

Proof. According to Lemma 4, the kth eigenvalue is

$$\lambda_k = (-1)^k 4\pi \int_\epsilon^R \exp(-2\pi r^2) L_k(4\pi r^2)\, r dr.$$

The following classical asymptotic expansion for Laguerre polynomials, valid for $x \in [\epsilon', R']$ with $0 < \epsilon' < R'$, will be essential [12, Thm. 8.22.2]:

$$\pi^{1/2} x^{1/4} k^{1/4} \exp\left(-\frac{x}{2}\right) L_k(x) = \cos\left(2(kx)^{1/2} - \frac{\pi}{4}\right)\left(1 + A_1(x)k^{-1/2} + \mathcal{O}(k^{-1})\right)$$
$$+ \sin\left(2(kx)^{1/2} - \frac{\pi}{4}\right)\left(B_1(x)k^{-1/2} + \mathcal{O}(k^{-1})\right).$$

Applying this formula with $x = 4\pi r^2$ yields

$$|\lambda_k| = k^{-1/4}I_0(k) + k^{-3/4}I_1(k) + \mathcal{O}(k^{-5/4}), \tag{6}$$

where

$$I_0(k) = 2^{3/2}\pi^{1/4}\int_\epsilon^R r^{1/2}\cos\left(4(k\pi)^{1/2}r - \frac{\pi}{4}\right)\,dr$$

and

$$\begin{aligned} I_1(k) =& 2^{3/2}\pi^{1/4}\int_\epsilon^R r^{1/2}A_1(4\pi r^2)\,\cos\left(4(k\pi)^{1/2}r - \frac{\pi}{4}\right)\,dr \\ &+ 2^{3/2}\pi^{1/4}\int_\epsilon^R r^{1/2}B_1(4\pi r^2)\,\sin\left(4(k\pi)^{1/2}r - \frac{\pi}{4}\right)\,dr. \end{aligned}$$

Consider the behavior of $I_0(k)$. Using a double-angle formula, we have

$$I_0(k) = 2^{3/2}\pi^{1/4}\int_\epsilon^R r^{1/2}\cos(4(k\pi)^{1/2}r)\,dr + 2^{3/2}\pi^{1/4}\int_\epsilon^R r^{1/2}\sin(4(k\pi)^{1/2}r)\,dr.$$

These are the Fourier cosine and sine integrals of the smooth function $r^{1/2}$ on the interval $[\epsilon, R]$, evaluated as $4(k\pi)^{1/2}$. As such, both integrals are $\mathcal{O}(k^{-1/2})$. A similar argument shows that the integrals in $I_1(k)$ are $\mathcal{O}(k^{-1/2})$ as well.

Finally, with a little more care one can show that

$$\limsup_{k\to\infty} k^{1/2}I_0(k) > 0.$$

In fact, write the function $r^{1/2} = \ell(r) + (r^{1/2} - \ell(r))$, where $\ell(r)$ is the linear function that interpolates between the endpoint values of $r^{1/2}$ at ϵ and R. The function $r^{1/2} - \ell(r)$ is a Lipschitz function on the real line with support in $[\epsilon, R]$. Consequently, the Fourier sine and cosine integrals of this function evaluated at $4(k\pi)^{1/2}$ are $\mathcal{O}(k^{-1})$. It is therefore enough to show that the integral

$$A_0(k) = \Re\left(e^{i\pi/4}\int_\epsilon^R \ell(y)\exp(4i(k\pi)^{1/2}r)\,dr\right)$$

has the property that

$$\limsup_{k\to\infty} k^{1/2}A_0(k) > 0.$$

This is easy to show directly.

For general domains, we obtain the weaker result that the sequence of eigenvalues is not absolutely summable.

PROPOSITION 11. *The series $\sum \lambda_k$ is not absolutely convergent.*

Proof. It is well known that $\{W(\phi_k, \phi_l)\}$ is an orthonormal basis for $L^2(\mathbb{R}^2)$. The equations $\iint_\Omega W(\phi_k, \phi_l) = \delta_{kl}\lambda_k$ imply that $\chi_\Omega = \sum \lambda_k W(\phi_k, \phi_k)$ in $L^2(\mathbb{R}^2)$. On the other hand, if $\sum |\lambda_k| < \infty$, then $\sum \lambda_k W(\phi_k, \phi_k)$ would have to converge uniformly to an element of $C_0(\mathbb{R}^2)$ (see Theorem 1, part (2)). This is clearly a contradiction.

We now examine the smoothness and decay of the eigenfunctions. We assume that Ω is an open set contained in a rectangle $[-B, B] \times [-T, T]$ of which all cross-sections in the ξ and t directions consist of at most M intervals. We now examine an equation of the form $L_{\chi_\Omega}\phi = \lambda\phi$, where λ is a nonzero real number and $\phi \in L^2(\mathbb{R})$. Fubini's theorem implies that

$$\lambda\phi(s) = \int K(s,t)\phi(t)\,dt$$

holds for all $s \in \mathbb{R}\backslash Z$, where Z is some set of measure zero. Lemma 5 yields the estimate

$$\begin{aligned} |\lambda\phi(s)| &\leq \frac{C}{1+|s|}\int_{-s-2T}^{-s+2T} |\phi(t)|\,dt \\ &\leq \frac{C\sqrt{4T}}{|s|+1}\|\phi\|_{L^2} \leq \frac{C'}{|s|+1} \end{aligned}$$

for all $s \in \mathbb{R}\backslash Z$. This last estimate implies that $\phi \in L^\infty(\mathbb{R})$. Now define

$$\mathcal{E}(s) = \sup_{|s'|\geq|s|} |\phi(s')|.$$

Note that $\mathcal{E}(s)$ is even and decreasing in $|s|$. The quantity $|\lambda\phi(s)|$ can now be estimated as follows:

$$\begin{aligned} |\lambda\phi(s)| &\leq \frac{C}{|s|+1}\int_{-s-2T}^{-s+2T} |\phi(t)|\,dt \\ &\leq \frac{4CT}{|s|+1}\mathcal{E}(|s|-2T) \end{aligned}$$

for all $s \in (\mathbb{R}\backslash Z) \cap \{s' : |s'| > 2T\}$. Therefore, given any $b > 1$, there is an s_0 such that if $|s| \geq |s_0|$ and $s \in \mathbb{R}\backslash Z$, then

$$\mathcal{E}(|s|+2T) \leq \frac{1}{b}\mathcal{E}(|s|).$$

Iterating this estimate yields

$$\mathcal{E}(|s|+2nT) \leq \frac{1}{b^n}\mathcal{E}(|s|).$$

It is straightforward to check that

$$|\phi(s)| \leq Cb^{-|s|/(2T)} \quad \forall s \in \mathbb{R}\backslash Z. \tag{7}$$

As a consequence, $\hat{\phi}$ is smooth and has all its derivatives in L^2.

Now, by Theorem 1, part (4), we have that

$$\iint_{\tilde{\Omega}} W(\hat{\phi}_k, \hat{\phi}_l) = \iint_{\Omega} W(\phi_k, \phi_l) = \lambda_k\delta_{kl},$$

where $\tilde{\Omega} = \{(\xi,t) : (-t,\xi) \in \Omega\}$. Hence $L_{\chi_{\tilde{\Omega}}}\hat{\phi}_k = \lambda_k\hat{\phi}_k$, and by the preceding discussion ϕ is smooth and has all its derivatives in L^2. Now for a given s_1,

$$\left.\frac{d^n\phi}{ds^n}\right|_{s=s_1} = \int \hat{\phi}(\xi)(2\pi i\xi)^n e^{2\pi i s_1\xi}d\xi.$$

Using the estimate analogous to equation (7) for $\hat{\phi}$ with $b >> 1$, one has $|\hat{\phi}(\xi)| \leq Ce^{-4\pi|\xi|}$. Then

$$\left|\frac{1}{n!}\frac{d^n\phi}{ds^n}\right|_{s=s1} \leq \frac{2C}{n!}\int_0^\infty e^{-4\pi\xi}(2\pi\xi)^n\, d\xi$$
$$\leq \frac{C}{n!}\frac{\Gamma(n)}{2^{n+1}} = \frac{C}{2^{n+1}n}.$$

Hence the power series of ϕ at $s = s_1$ converges in some interval around s_1. Because of the symmetry in the role of ϕ and $\hat{\phi}$, the same observations hold for $\hat{\phi}$. The preceding discussion is summarized in the following theorem.

THEOREM 12. *Suppose Ω is an open set contained in the interval $[-B,B] \times [-T,T]$ with the property that all its cross-sections in both the ξ and t directions consist of at most M intervals. Then*

(1) *for any $b > 0$, there is a constant C_b such that*

$$|\phi(s)| \leq C_b e^{-b|s|} \quad \forall s \quad \text{and} \quad |\hat{\phi}(\xi)| \leq C_b e^{-b|\xi|} \quad \forall \xi,$$

and

(2) *ϕ and $\hat{\phi}$ are analytic and have all their derivatives in L^2.*

Note that (1) actually implies (2) in Theorem 12 by the Paley–Wiener theorems [13], although we have chosen to give the elementary argument.

It is instructive to consider the case when Ω is a ball centered at the origin. It is well known that, in this case, the ϕ_k are Hermite functions (see Lemma 4). The Hermite functions $h_j(t)$ in Lemma 4 will satisfy estimates of the form $h_j(t) \leq Ce^{-(\pi-\epsilon)t^2}$ for any $\epsilon > 0$. It is directly evident that they will then satisfy the weaker inequalities in Theorem 12. This theorem states that this weaker decay statement holds for general regions in the time-frequency plane. We do not know whether these estimates can be improved.

4. Numerical examples. For illustration purposes, we provide four numerical examples of time-frequency localization. These examples are obtained by discretizing the kernel in the integral representation of the operator given in Lemma 5. We consider localization on domains of the form of a "zigzag," a disk, a rectangle, and a parallelogram, as indicated in Fig. 1. Note that while the operators here are all Hermitian, so that they have real eigenvalues, they do not generally have real eigenfunctions unless their kernels are real. This is true if the symbol $S(\xi,t)$ satisfies $S(-\xi,t) = S(\xi,t)$. For clarity of eigenfunction representation, projection domains were chosen to allow this symmetry when possible.

Localization on the unusual zigzag domain (Fig. 1a) produces the unfamiliar eigenfunctions in Fig. 2 (as these are complex, we have plotted their magnitude). Eight samples per unit length are used, over the time domain $[-8,8]$. Most of the energy is concentrated in the desired region $[-2,2]$, although there is some leakage. Note that the first three eigenfunctions have one, two, and three peaks, respectively. For the case of the disk (Fig. 1b), a plot of the eigenvalues is shown in Fig. 3. As noted in Lemma 4, the eigenfunctions in this case are the well-known hermite functions. Next, localization on the rectangle (Fig. 1c) allows comparison to the prolate spheroidal wavefunctions. The first nine solutions for the prolate spheroidal and Weyl operator cases are depicted in Figs. 4a and b. Again, eight samples per unit length are used over the time domain $[-8,8]$. While there is energy leakage outside the desired domain

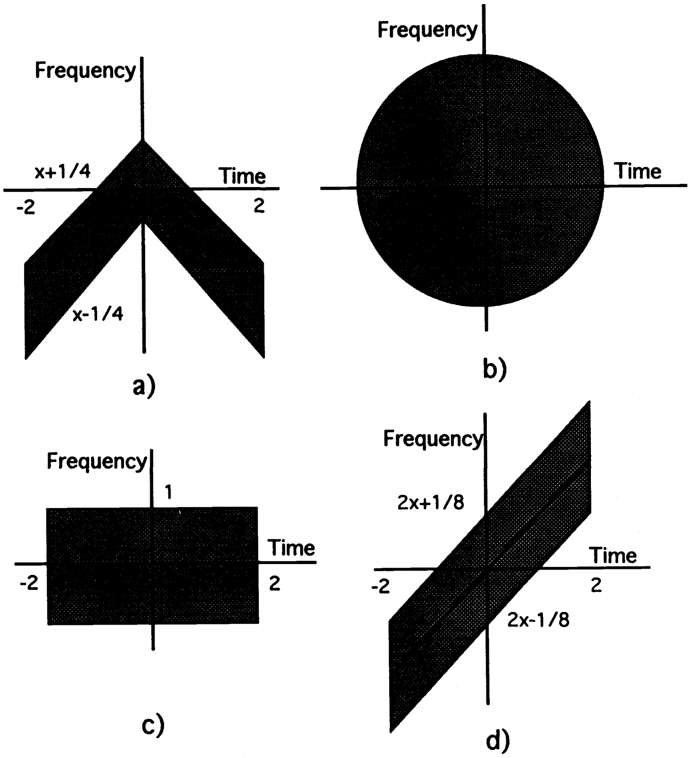


FIG. 1. *Localization Domains:* (a) *zigzag domain*; (b) *disk of area* 50; (c) 4 × 2 *rectangle*; (d) *parallelogram.*

$[-2, 2]$, the amount differs in the two cases. The prolate spheroidal wavefunctions have $1/x$ decay, while the Weyl eigenfunctions have exponential decay. This difference is visible when one compares the last three eigenfunctions in each case. A plot of the eigenvalues is also included.

Finally, we provide a simple illustration of these ideas in the context of filtering Gaussian noise from a corrupted linear FM (chirp) signal. Figures 5a and 6a show, respectively, the Wigner distribution intensity plot and the actual plot of the real part of a linear FM chirp centered at zero frequency. (Although chirp signals used in radar are not centered at zero frequency, that is immaterial for our purposes because of the covariance of the Wigner distribution under time and frequency shifts (Theorem 1, part (5).) In the discretization, 16 samples per unit length are used, over the time domain $[-2, 2]$. No windowing is applied to remove the echo effect in these Wigner plots, although we have found a simple cosine-squared window to be effective. Figures 5b and 6b show this signal after 0 dB Gaussian white noise has been added. To filter the noise, we note that theoretically, the Wigner distribution of a chirp signal is a measure concentrated along a diagonal line corresponding to the slope of the chirp. In particular, a chirp can be localized in any domain containing its time-frequency support. As an elementary example of time-frequency localization, the noisy chirp of Fig. 6b is projected onto the first two eigenfunctions (weighted according to the eigenvalues) for the domain in Fig. 1d. This results in the signal in Fig. 6c, with a Wigner distribution as shown in Fig. 5c. A plot of the eigenvalues for this localization operator is provided in Fig. 6d. In fact, the chirp in Fig. 6a is orthogonal to the second eigenfunction, illustrating an interesting fact. Numerically, the chirp appears to be an exact solution to the problem of localizing onto an infinite diagonal band domain. For

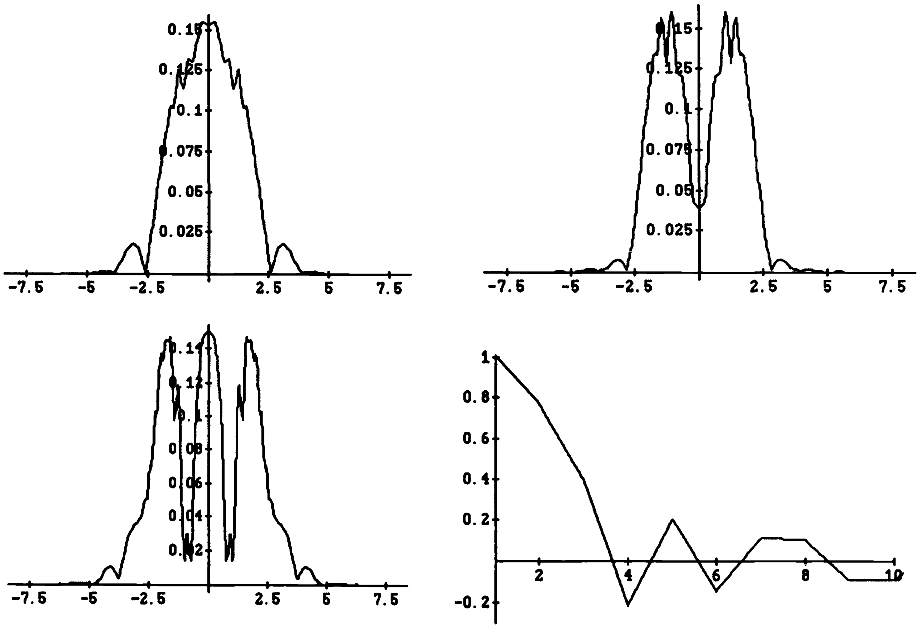


FIG. 2. *Magnitude of eigenfunctions* 1–3 *for localization onto a zigzag domain* (*Fig.* 1a), *with a plot of eigenvalues sorted according to decreasing magnitude.*

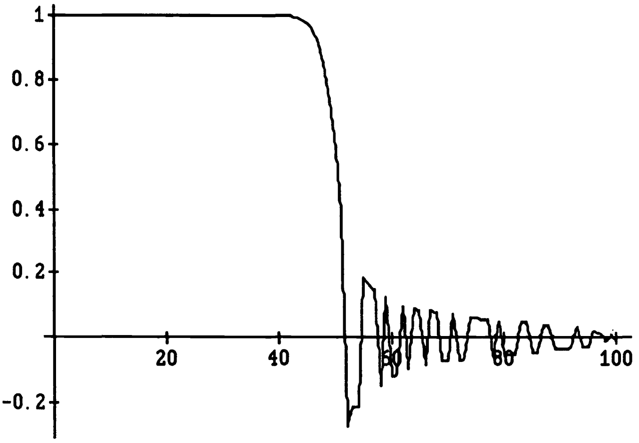


FIG. 3. *Weyl eignevalue plot for localization onto a disk of area* 50 (*Fig.* 1b), *sorted according to decreasing magnitude.*

more numerical examples of time-frequency localization, the reader is referred to [2], [3].

It is interesting at this point to remark on a conjecture of Flandrin [10]: for localization onto convex domains, the top eigenvalue is bounded above by 1. This seems to hold (at least numerically) in our examples, even for Fig. 1a, which is not convex. However, something like convexity is certainly necessary in general, since we have nonconvex numerical examples where the top eigenvalue exceeds 4.

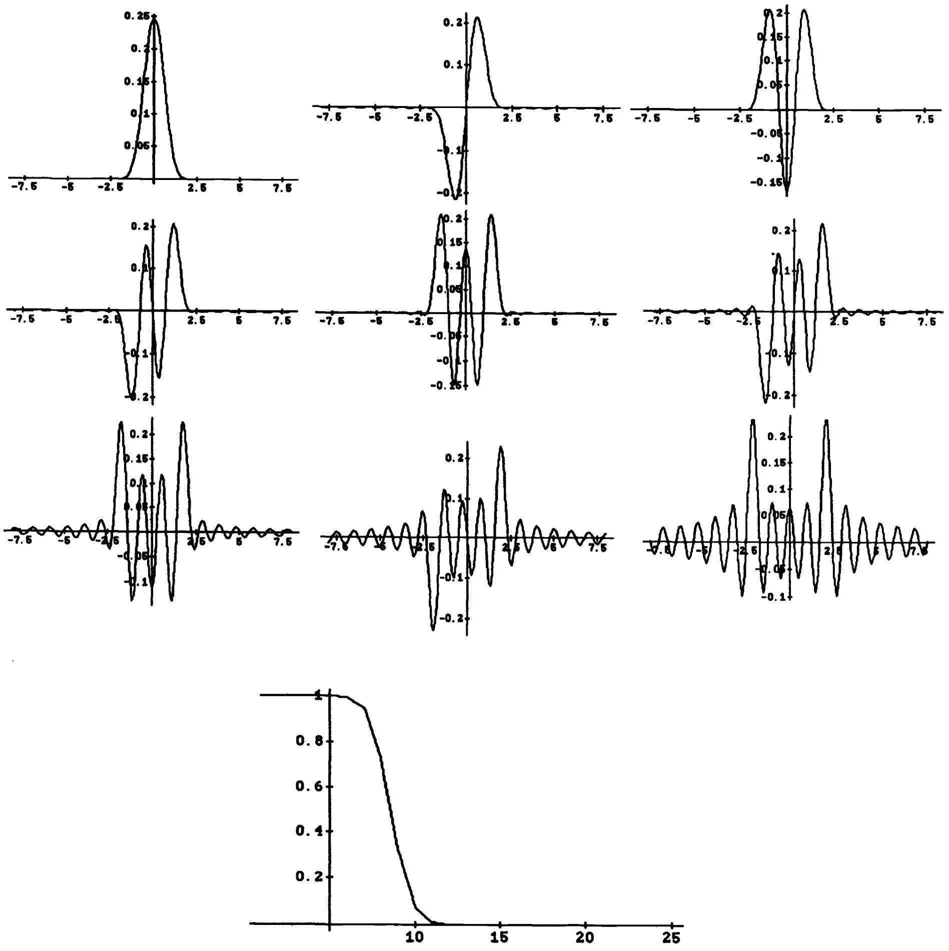


FIG. 4a. *Prolate spheroidical wavefunctions* 1–9 *for localization on a rectangle* (*Fig.* 1c), *with eigenvalue plot.*

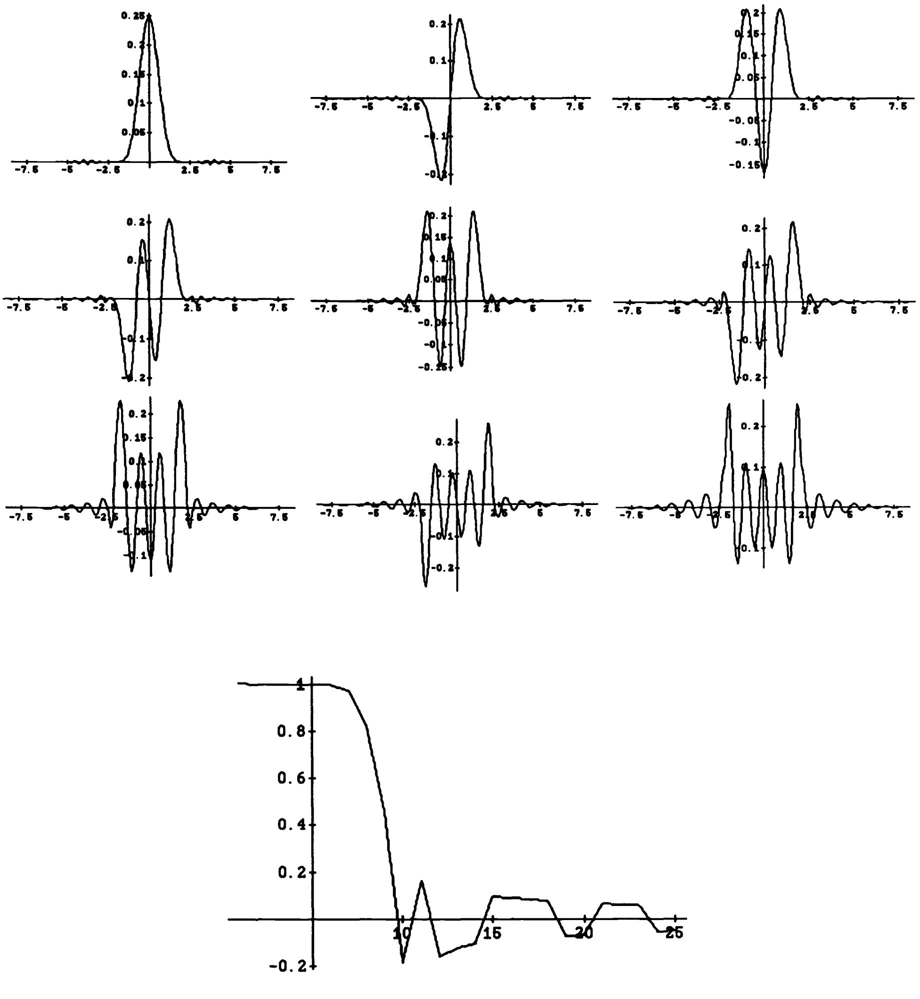


FIG. 4b. *Weyl eigenfunctions* 1–9 *for localization on a rectangle* (*Fig.* 1c), *with a plot of eigenvalues in decreasing magnitude.*

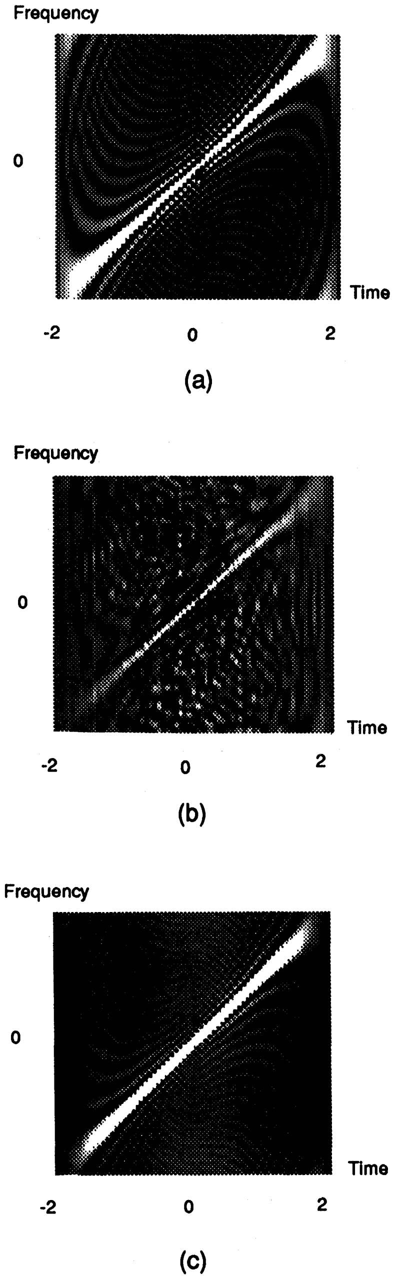


FIG. 5. *Wigner distribution of* (a) *chirp signal*; (b) *chirp in* 0 dB *Gaussian white noise*; (c) *filtered noisy signal.*

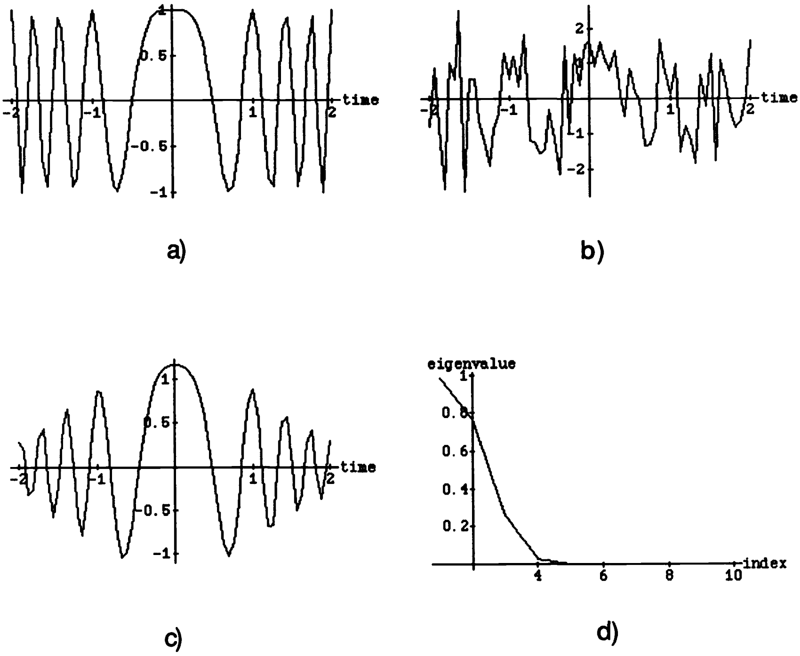


FIG. 6. *Plots of the real part of* (a) *chirp signal,* (b) *chirp signal in* 0 dB *Gaussian white noise, and* (c) *filtered noisy signal.* (d) *is an eigenvalue plot. The first two eigenfunctions, weighted by their eigenvalues, are used in the filtering.*

Acknowledgments. We are grateful to the referees for bringing references [10]–[12] to our attention, and for making several other valuable suggestions. We especially thank the referee who suggested a result along the lines of Proposition 10.

REFERENCES

[1] G. B. FOLLAND, *Harmonic Analysis in Phase Space*, Princeton University Press, Princeton, NJ, 1989.

[2] F. HLAWATSCH, W. KOZEK, AND W. KRATTENTHALER, *Time frequency subspaces and their application to time-varying filtering*, Proc. International Conference on Acoustics, Speech, and Signal Processing, 1990, pp. 1609–1610.

[3] F. HLAWATSCH AND W. KOZEK, *Time-frequency analysis of linear signal subspaces*, Proc. International Conference on Acoustics, Speech, and Signal Processing, 1991, pp. 2045–2048.

[4] H. WEYL, *Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung)*, Math. Ann., 71 (1911), pp. 441–479.

[5] I. DAUBECHIES, *Time-frequency localization operators—a geometric phase space approach,* I, IEEE Trans. Inform. Theory, 34 (1988), pp. 605–612.

[6] I. DAUBECHIES AND T. PAUL, *Time-frequency localization operators—a geometric phase space approach,* II, Inverse Problems, 4 (1988), pp. 661–680.

[7] D. SLEPIAN AND H. O. POLLACK, *Prolate spheroidal wavefunctions, fourier analysis and uncertainty*: I, Bell Syst. Tech. J., 40 (1961), pp. 43–64.

[8] H. J. Landau and H. O. Pollack, *Prolate spheroidal wavefunctions, fourier analysis and uncertainty*: II, III, Bell Syst. Tech. J., 40,41 (1961,1962), pp. 43–64, pp. 1295–1336.

[9] J. Pool, *Mathematical aspects of the Weyl correspondence*, J. Math. Phys., 7 (1966), pp. 66–76.

[10] P. Flandrin, *Maximal signal energy concentration in a time-frequency domain*, Proc. International Conference on Acoustics, Speech, and Signal Processing, 1988, pp. 2176–2179.

[11] A. J. E. M. Janssen, *Positivity of weighted Wigner distributions*, SIAM J. Math. Anal., 12 (1981), pp. 752–758.

[12] G. Szego, *Orthogonal polynomials*, Vol. 23, Amer. Math. Soc. Colloq. Publ., Providence, RI, 1959.

[13] Y. Katznelson, *An Introduction to Harmonic Analysis*, John Wiley, New York, 1968.

SIAM J. MATH. ANAL.
Vol. 24, No. 6, pp. 1395–1413, November 1993

SPATIAL DECAY ESTIMATES IN TIME-DEPENDENT STOKES FLOW*

K. A. AMES†, L. E. PAYNE‡, AND P. W. SCHAEFER§

Abstract. This paper considers the time-dependent slow flow of an incompressible viscous fluid in a semi-infinite cylindrical pipe of smooth cross section. An exponential decay estimate in terms of the distance from the finite end of the pipe is obtained from a second-order differential inequality for a weighted energy integral defined on the solutions of the system. The decay constant depends only on the geometry and the first positive eigenvalues for the fixed and free membrane problems for the cross sectional geometry. The paper also indicates how to bound the total weighted energy.

Key words. Stokes flow, decay estimates

AMS subject classifications. 35Q30, 35B45, 76D07

1. Introduction.

Although principles of Saint-Venant type, estimating the spatial decay of solutions of various elliptic boundary value problems have been extensively investigated in the literature of the past century (see [5], [8], and the references therein), the study of spatial decay of solutions of time-dependent problems is of relatively recent origin. The first work in this area appears to be that of Boley [3] in 1958. For subsequent investigations on the spatial decay of solutions of parabolic equations, also see the references cited in [5] and [8].

In this paper, we investigate the slow flow of an incompressible viscous fluid in a semi-infinite cylindrical pipe. At the finite end of the pipe, a time-dependent velocity profile is prescribed, adherence is assumed on the lateral surface, and the fluid is assumed to be initially at rest. We derive an explicit inequality which implies exponential decay of a weighted energy expression as a function of the distance from the finite end of the pipe. Of course, the solution will not exhibit spatial decay for each t unless the net flow through the finite end of the pipe is zero for each t. The case of nonzero net entry flow will be considered in a later paper.

The two-dimensional version of the pipe flow problem has been studied by Lin [9]. In this case, he was able to eliminate the troublesome pressure term by introducing a stream function. This feature does not carry over to the pipe flow problem here, so different techniques must be developed for establishing the decay estimate.

It should be remarked that Elcrat and Sigillito [4] actually looked at the question of spatial decay for the dynamical Navier–Stokes equations, but their method required an assumption on an auxiliary function that is not generally satisfied. For decay results in steady pipe flow, see the references cited in Horgan and Knowles [5] and the paper of Ames and Payne [1].

The outline for this paper is as follows. We formulate the boundary value problem, which describes transient Stokes flow in three space, and define the energy function

* Received by the editors October 8, 1992; accepted for publication (in revised form) February 24, 1993.

† Department of Mathematical Sciences, University of Alabama in Huntsville, Huntsville, Alabama 35899.

‡ Department of Mathematics, Cornell University, Ithaca, New York 14853. The work of this author was supported by National Science Foundation grant DMS 9100786.

§ Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37996. The work of this author was supported by a University of Tennessee Faculty Development Grant.

in §2. After recording in §3 some auxiliary results that we shall use, we determine a second-order differential inequality for the energy expression in §§4 and 5. The exponential decay inequality which follows from the differential inequality is given in (5.34). In §6, we estimate the total weighted energy by techniques developed in the body of this paper, and in §7, we make some concluding remarks.

2. Statement of the problem. Let R denote the interior of a semi-infinite cylindrical pipe of an arbitrary, smooth cross section with generator parallel to the x_3-axis. The end (entrance) of the pipe in the $x_3 = 0$ plane is denoted by D_0 and comprises part of the boundary ∂R of R. We let

$$R_z = \{(x_1, x_2, x_3) \mid (x_1, x_2) \in D_0,\ x_3 > z \geq 0\}$$

denote the subdomain of R for which $x_3 > z$ and let

$$D_z = \{(x_1, x_2, x_3) \mid (x_1, x_2) \in D_0,\ x_3 = z\}$$

denote the part of ∂R_z in the plane $x_3 = z \geq 0$.

The velocity field $u_i(x_1, x_2, x_3, t)$ for $i = 1, 2, 3$ and the pressure $p(x_1, x_2, x_3, t)$ for the time-dependent Stokes flow of an incompressible viscous fluid are assumed to be classical solutions of the initial-boundary value problem:

$$u_{i,t} = \nu \Delta u_i - p_{,i} \quad \text{in } R \times (0, \infty), \tag{2.1}$$

$$u_{i,i} = 0 \quad \text{in } \overline{R} \times (0, \infty), \tag{2.2}$$

$$u_i = 0 \quad \text{on } \partial R \setminus D_0 \times (0, \infty), \tag{2.3}$$

$$u_i = f_i(x_1, x_2, t) \quad \text{in } \overline{D}_0 \times (0, \infty), \tag{2.4}$$

$$u_i = 0 \quad \text{in } R \times \{0\}, \tag{2.5}$$

$$u_i, u_{i,j}, u_{i,t}, p = o\left(x_3^{-1}\right) \quad \text{uniformly in } x_1, x_2, t \quad \text{as } x_3 \to \infty, \tag{2.6}$$

where Δ is the three space Laplace operator, ν is the constant kinematic viscosity, and the comma (partial differentiation) and repeated index (summation) conventions are used. In this work, Latin subscripts range over 1, 2, 3 while Greek subscripts range over 1, 2 unless noted otherwise. We assume that the prescribed functions (entrance profile) f_i are continuously differentiable and vanish on $\partial D_0 \times [0, \infty)$ and that $\nu = 1$ without loss of generality since we can rescale the time variable.

We define a weighted energy integral for solutions u_i of (2.1)–(2.6) by (no summation on τ)

$$E(z, t) = \int_0^t \int_{R_z} (\xi - z)\left[u_{i,j} u_{i,j} + k u_{i,\tau} u_{i,\tau}\right] dx d\tau \tag{2.7}$$

and note that

$$\frac{\partial E}{\partial z} = -\int_0^t \int_{R_z} \left[u_{i,j} u_{i,j} + k u_{i,\tau} u_{i,\tau}\right] dx d\tau, \tag{2.8}$$

$$\frac{\partial^2 E}{\partial z^2} = \int_0^t \int_{D_z} \left[u_{i,j} u_{i,j} + k u_{i,\tau} u_{i,\tau}\right] dA d\tau, \tag{2.9}$$

where k is a parameter to be chosen. Our aim is to obtain a second-order differential inequality for E from which we can deduce an exponential decay estimate of the form

$$E(z, t) \leq E(0, t) e^{-\sigma z}, \qquad \sigma > 0, \tag{2.10}$$

where $0 \leq z < \infty$.

3. Auxiliary inequalities. We shall make frequent use of the Schwarz inequality and the arithmetic mean-geometric mean (AG) inequality in our derivation of the second-order differential inequality we seek. In addition, we need the following three inequalities.

Let S be a plane domain with sufficiently smooth boundary ∂S, and let v be a sufficiently smooth function defined on the closure $\overline{S}$ of S.

1. If $v = 0$ on ∂S, then

$$\lambda_1 \int_S v^2 dA \leq \int_S v_{,\alpha} v_{,\alpha} dA, \tag{3.1}$$

where λ_1 is the smallest positive eigenvalue of

$$\Delta w + \lambda w = 0 \text{ in } S, \quad w = 0 \quad \text{on } \partial S.$$

2. If $\partial v / \partial n = 0$ on ∂S and $\int_S v dA = 0$, then

$$\mu_2 \int_S v^2 dA \leq \int_S v_{,\alpha} v_{,\alpha} dA, \tag{3.2}$$

where μ_2 is the smallest positive eigenvalue of

$$\Delta w + \mu w = 0 \quad \text{in } S, \quad \frac{\partial w}{\partial n} = 0 \quad \text{on } \partial S, \quad \int_S w dA = 0.$$

3. If g is a continuously differentiable function on $\overline{S}$ and $\int_S g dA = 0$, then there exists a vector function v_α such that

$$v_{\alpha,\alpha} = g \quad \text{in } S, \quad v_\alpha = 0 \quad \text{on } \partial S, \tag{3.3}$$

and a positive constant C depending only on the geometry of S such that

$$\int_S v_{\alpha,\beta} v_{\alpha,\beta} dA \leq C \int_S (v_{\alpha,\alpha})^2 dA. \tag{3.4}$$

The first two inequalities are the result of the variational characterization of the smallest positive eigenvalue of the respective eigenvalue problem, whereas the third inequality is established in [2]. The latter inequality appears to have first been used in [7]. The third implication above asserts the existence of a vector function v_α which is, in fact, not unique. We require only the existence of such a vector function in our derivation and not an explicit solution. We refer the reader to [1] for a brief discussion about the constant C and to [6] for an explicit upper bound for the optimal C when S is a star-shaped domain—an assumption we will make about the domain D_0 (see Lemma 3 in §5).

4. Energy estimation—first part. We are now ready to derive the desired second-order differential inequality for E. Our derivation consists of two parts. In this section, we bound the more easily estimated terms and in the next section, we bound a term complicated by the presence of the pressure function multiplied by the time derivative of u_3 and then determine the exponential decay estimate for the energy expression E.

We consider the energy integral

$$E(z,t) = \int_0^t \int_{R_z} (\xi - z) \left[u_{i,j} u_{i,j} + k u_{i,\tau} u_{i,\tau} \right] dx d\tau, \tag{4.1}$$

where we recall that there is no summation on the τ subscript and k is a parameter to be chosen. On integrating by parts and using (2.1), we have

$$\begin{aligned} E(z,t) = &-\int_0^t \int_{R_z} u_i u_{i,3} dx d\tau - \int_0^t \int_{R_z} (\xi - z) u_i \left[u_{i,\tau} + p_{,i}\right] dx d\tau \\ &+ k \int_0^t \int_{R_z} (\xi - z) u_{i,\tau} \left[\Delta u_i - p_{,i}\right] dx d\tau. \end{aligned}$$

We integrate by parts again and drop two spatial integral terms which are negative to obtain the inequality

$$\begin{aligned} E(z,t) \leq &-\int_0^t \int_{R_z} u_i u_{i,3} dx d\tau - k \int_0^t \int_{R_z} u_{i,\tau} u_{i,3} dx d\tau \\ &+ \int_0^t \int_{R_z} u_3 p dx d\tau + k \int_0^t \int_{R_z} u_{3,\tau} p dx d\tau \\ = &\, I_1 + I_2 + I_3 + I_4. \end{aligned} \tag{4.2}$$

We now proceed to estimate the first three integral terms.

We use the Schwarz inequality, (3.1), and the AG inequality to obtain

$$\begin{aligned} I_1 &\leq \int_0^t \int_z^\infty \left(\int_{D_\xi} u_i u_i dA \int_{D_\xi} u_{i,3} u_{i,3} dA \right)^{\frac{1}{2}} d\xi d\tau \\ &\leq \int_0^t \int_z^\infty \left(\frac{1}{\lambda_1} \int_{D_\xi} u_{i,\alpha} u_{i,\alpha} dA \int_{D_\xi} u_{i,3} u_{i,3} dA \right)^{\frac{1}{2}} d\xi d\tau \\ &\leq \frac{1}{2\sqrt{\lambda_1}} \left\{ \int_0^t \int_{R_z} u_{i,\alpha} u_{i,\alpha} dx d\tau + \int_0^t \int_{R_z} u_{i,3} u_{i,3} dx d\tau \right\}. \end{aligned}$$

Thus we can write

$$I_1 \leq \frac{1}{2\sqrt{\lambda_1}} \left(-\frac{\partial E}{\partial z} \right). \tag{4.3}$$

In a similar manner, it follows that

$$I_2 \leq \frac{k}{2} \left\{ \sigma_1 \int_0^t \int_{R_z} u_{i,\tau} u_{i,\tau} dx d\tau + \frac{1}{\sigma_1} \int_0^t \int_{R_z} u_{i,3} u_{i,3} dx d\tau \right\}.$$

If we let $k = 1/\lambda_1$ and choose $\sigma_1 = 1/\sqrt{\lambda_1}$, then we obtain

$$I_2 \leq \frac{1}{2\sqrt{\lambda_1}} \left(-\frac{\partial E}{\partial z} \right). \tag{4.4}$$

Due to the presence of the pressure function in the I_3 integral, we proceed in a different manner to bound this term. We first note that for any $z^* > 0$, by (2.2) and (2.3),

$$\begin{aligned}\int_{D_z} u_3 dA &= \int_{D_{z^*}} u_3 dA - \int_z^{z^*}\int_{D_\xi} u_{3,3} dA d\xi \\ &= \int_{D_{z^*}} u_3 dA + \int_z^{z^*}\int_{D_\xi} u_{\alpha,\alpha} dA d\xi \\ &= \int_{D_{z^*}} u_3 dA,\end{aligned}$$

and thus the area mean value of u_3 is the same over each section. Since u_3 is assumed to vanish at infinity, we conclude that this value is zero for all $z \geq 0$ and hence requires that f_3 satisfy

$$\int_{D_0} f_3 dA = 0 \quad \text{for all } t \geq 0. \tag{4.5}$$

Under this assumption, there exists a vector function (see (3.3)) ω_α which satisfies

$$\omega_{\alpha,\alpha} = u_3 \quad \text{in } D_\xi, \quad \omega_\alpha = 0 \quad \text{on } \partial D_\xi,$$

for each $\xi \geq 0$ and for which (3.4) holds.

Using the vector function ω_α in I_3, it follows that

$$\begin{aligned}I_3 &= \int_0^t \int_{R_z} \omega_{\alpha,\alpha} p dx d\tau \\ &= -\int_0^t \int_{R_z} \omega_\alpha p_{,\alpha} dx d\tau \\ &= \int_0^t \int_{R_z} \omega_\alpha \left[u_{\alpha,\tau} - \Delta u_\alpha\right] dx d\tau \\ &= \int_0^t \int_{R_z} \omega_\alpha u_{\alpha,\tau} dx d\tau + \int_0^t \int_{D_z} \omega_\alpha u_{\alpha,3} dA d\tau + \int_0^t \int_{R_z} \omega_{\alpha,i} u_{\alpha,i} dx d\tau.\end{aligned}$$

Now by means of Schwarz's inequality, (3.1), and the weighted AG inequality, we have

$$\begin{aligned}I_3 \leq& \frac{1}{2\sqrt{\lambda_1}}\left\{\sigma_2 \int_0^t \int_{R_z} u_{\alpha,\tau} u_{\alpha,\tau} dx d\tau + \frac{1}{\sigma_2}\int_0^t \int_{R_z} \omega_{\alpha,\beta}\omega_{\alpha,\beta} dx d\tau\right\} \\ &+ \frac{1}{2}\left\{\sigma_3 \int_0^t \int_{R_z} u_{\alpha,3} u_{\alpha,3} dx d\tau + \frac{1}{\sigma_3}\int_0^t \int_{R_z} \omega_{\alpha,3}\omega_{\alpha,3} dx d\tau\right\} \\ &+ \frac{1}{2}\left\{\sigma_4 \int_0^t \int_{R_z} u_{\alpha,\beta} u_{\alpha,\beta} dx d\tau + \frac{1}{\sigma_4}\int_0^t \int_{R_z} \omega_{\alpha,\beta}\omega_{\alpha,\beta} dx d\tau\right\} \\ &+ \frac{1}{2\sqrt{\lambda_1}}\left\{\sigma_5 \int_0^t \int_{D_z} u_{\alpha,3} u_{\alpha,3} dA d\tau + \frac{1}{\sigma_5}\int_0^t \int_{D_z} \omega_{\alpha,\beta}\omega_{\alpha,\beta} dA d\tau\right\}.\end{aligned} \tag{4.6}$$

The integrals of the auxiliary function ω_α can be bounded in terms of u_3 by means of (3.1) and (3.4). Thus we can write

$$\begin{aligned}\int_0^t \int_{R_z} \omega_{\alpha,\beta}\omega_{\alpha,\beta} dx d\tau &\leq C \int_0^t \int_{R_z} (u_3)^2 dx d\tau \\ &\leq \frac{C}{\lambda_1}\int_0^t \int_{R_z} u_{3,\beta} u_{3,\beta} dx d\tau,\end{aligned}$$

$$\begin{aligned}\int_0^t\int_{R_z}\omega_{\alpha,3}\omega_{\alpha,3}dxd\tau &\le \frac{1}{\lambda_1}\int_0^t\int_{R_z}\omega_{\alpha,\beta 3}\omega_{\alpha,\beta 3}dxd\tau\\ &\le \frac{C}{\lambda_1}\int_0^t\int_{R_z}(u_{3,3})^2dxd\tau.\end{aligned}$$

We now substitute these estimates into (4.6) and choose

$$\sigma_2=\sqrt{C}/\lambda_1\,,\qquad \sigma_3=\sigma_4=\sigma_5=\sqrt{C}\Big/\sqrt{\lambda_1}\,,$$

so that

$$\begin{aligned}I_3 \le &\frac{1}{2}\sqrt{\frac{C}{\lambda_1}}\left\{\int_0^t\int_{R_z}ku_{\alpha,\tau}u_{\alpha,\tau}dxd\tau+\int_0^t\int_{R_z}u_{3,\beta}u_{3,\beta}dxd\tau\right\}\\ &+\frac{1}{2}\sqrt{\frac{C}{\lambda_1}}\int_0^t\int_{R_z}u_{i,j}u_{i,j}dxd\tau\\ &+\frac{\sqrt{C}}{2\lambda_1}\left\{\int_0^t\int_{D_z}u_{\alpha,3}u_{\alpha,3}dAd\tau+\int_0^t\int_{D_z}u_{3,\beta}u_{3,\beta}dAd\tau\right\}.\end{aligned}$$

Finally, for I_3 we have the estimate

$$I_3\le\sqrt{\frac{C}{\lambda_1}}\left(-\frac{\partial E}{\partial z}\right)+\frac{\sqrt{C}}{2\lambda_1}\frac{\partial^2E}{\partial z^2}. \tag{4.7}$$

The result of estimating the first three integrals in (4.2) is

$$I_1+I_2+I_3\le\frac{1+\sqrt{C}}{\sqrt{\lambda_1}}\left(-\frac{\partial E}{\partial z}\right)+\frac{\sqrt{C}}{2\lambda_1}\frac{\partial^2E}{\partial z^2}. \tag{4.8}$$

We now seek an appropriate estimate for the integral I_4. As this involves a much more complicated calculation, we compute the I_4 bound in the next section.

5. Energy estimation—second part. We need to bound I_4 in terms of $\partial E/\partial z$ and $\partial^2E\big/\partial z^2$ in order to derive the second-order differential inequality for the energy expression E. To accomplish this, we introduce two auxiliary problems for which only the existence of a solution will be required.

Let φ be a solution of the boundary value problem

$$\Delta\varphi=u_{3,t}\quad\text{in } R_z,\qquad \frac{\partial\varphi}{\partial n}=0\quad\text{on }\partial R_z. \tag{5.1}$$

It is clear that

$$\int_{R_z}u_{3,t}dx=\int_z^\infty\int_{D_\xi}u_{3,t}dAd\xi=0, \tag{5.2}$$

and since $u_{3,t}\to 0$ as $x_3\to\infty$, it follows from the Phragmèn–Lindelöf theory that there exists a solution φ which vanishes together with its spatial derivatives as $x_3\to\infty$.

By means of φ, we can write

$$\begin{aligned} I_4 &= k\int_0^t\int_{R_z} p\Delta\varphi dx d\tau = -k\int_0^t\int_{R_z} \varphi_{,i}p_{,i}dx d\tau \\ &= k\int_0^t\int_{R_z} \varphi_{,i}\left[u_{i,\tau} - \left(u_{i,j}-u_{j,i}\right)_{,j}\right]dx d\tau \\ &= k\int_0^t\int_{R_z} \varphi_{,i}u_{i,\tau}dx d\tau + k\int_0^t\int_{D_z}\varphi_{,\alpha}\left(u_{\alpha,3}-u_{3,\alpha}\right)dA d\tau \\ &\quad - k\int_0^t\int_z^\infty\int_{\partial D_\xi}\varphi_{,i}\left(u_{i,j}-u_{j,i}\right)n_j ds d\xi d\tau, \end{aligned}$$

and by Schwarz's inequality, we have
(5.3)

$$\begin{aligned} I_4 \leq & k\left(\int_0^t\int_{R_z}\varphi_{,i}\varphi_{,i}dx d\tau\right)^{\frac{1}{2}}\left(\int_0^t\int_{R_z}u_{i,\tau}u_{i,\tau}dx d\tau\right)^{\frac{1}{2}} \\ &+ k\left(\int_0^t\int_{D_z}\varphi_{,\alpha}\varphi_{,\alpha}dA d\tau\right)^{\frac{1}{2}}\left(\int_0^t\int_{D_z}\left(u_{\alpha,3}-u_{3,\alpha}\right)\left(u_{\alpha,3}-u_{3,\alpha}\right)dA d\tau\right)^{\frac{1}{2}} \\ &+ k\left(\int_0^t\int_z^\infty\int_{\partial D_\xi}\left|\text{grad}_s\varphi\right|^2 ds d\xi d\tau\right)^{\frac{1}{2}}\left(\int_0^t\int_z^\infty\int_{\partial D_\xi}\left(u_{i,j}-u_{j,i}\right)u_{i,j}ds d\xi d\tau\right)^{\frac{1}{2}} \end{aligned}$$

since

$$\varphi_{,i}\left(u_{i,j}-u_{j,i}\right)n_j = \tfrac{1}{2}\left(\varphi_{,i}n_j - \varphi_{,j}n_i\right)\left(u_{i,j}-u_{j,i}\right).$$

In (5.3), $\text{grad}_s\varphi$ denotes the tangential component of the gradient of φ.

We must now determine suitable estimates for the integrals of φ in (5.3). We develop these bounds in the following lemmas. First, we deduce the following.

LEMMA 1. *If $\int_{D_{x_3}}\varphi dA = 0$ as $x_3 \to \infty$, then*

$$\int_{D_z}\varphi dA = 0 \quad \textit{for all } z \geq 0. \tag{5.4}$$

Proof. By (5.1), we have

$$0 = \int_{D_{x_3}} u_{3,t}dA = \int_{D_{x_3}}\varphi_{,ii}dA = \int_{D_{x_3}}\varphi_{,33}dA = \left(\int_{D_{x_3}}\varphi dA\right)_{,33},$$

which implies that

$$\int_{D_{x_3}}\varphi dA = ax_3 + b.$$

Hence by hypothesis, the conclusion follows.

LEMMA 2. *If φ, $\varphi_{,i} \to 0$ uniformly as $x_3 \to \infty$, then*

$$\int_{R_z}\varphi_{,i}\varphi_{,i}dx \leq \frac{1}{\mu_2}\int_{R_z}(u_{3,t})^2 dx, \tag{5.5}$$

$$\int_{D_z}\varphi_{,\alpha}\varphi_{,\alpha}dA \leq \frac{2}{\sqrt{\mu_2}}\int_{R_z}(u_{3,t})^2 dx. \tag{5.6}$$

Proof. For (5.5), we have

$$\begin{aligned}\int_{R_z} \varphi_{,i}\varphi_{,i}dx &= \int_{R_z} \varphi\Delta\varphi dx \\ &= \int_{R_z} \varphi u_{3,t}dx \\ &\leq \left(\int_{R_z} \varphi^2 dx\right)^{\frac{1}{2}} \left(\int_{R_z} (u_{3,t})^2 dx\right)^{\frac{1}{2}} \\ &\leq \left(\frac{1}{\mu_2}\int_{R_z} \varphi_{,i}\varphi_{,i}dx\right)^{\frac{1}{2}} \left(\int_{R_z} (u_{3,t})^2 dx\right)^{\frac{1}{2}},\end{aligned}$$

from which the result follows. We note that we used (5.4) in applying inequality (3.2).

We establish (5.6) by observing that

$$\int_{R_z} \varphi_{,3}\left[\Delta\varphi - u_{3,t}\right] dx = 0$$

implies the identity

$$\frac{1}{2}\int_{D_z} \varphi_{,i}\varphi_{,i}dA = \int_{R_z} \varphi_{,3}u_{3,t}dx.$$

Then by Schwarz's inequality, we can write

$$\frac{1}{2}\int_{D_z} \varphi_{,\alpha}\varphi_{,\alpha}dA \leq \left(\int_{R_z} \varphi_{,i}\varphi_{,i}dx\right)^{\frac{1}{2}} \left(\int_{R_z} (u_{3,t})^2 dx\right)^{\frac{1}{2}},$$

and by applying (5.5), we obtain (5.6).

LEMMA 3. *If φ, $\varphi_{,i} \to 0$ uniformly as $x_3 \to \infty$ and if D_z is star-shaped with respect to a point (origin) in D_z, then*

$$\int_z^\infty \int_{\partial D_\xi} |\text{grad}_s\varphi|^2 ds d\xi \leq \frac{2}{h_0}\left[\frac{1}{\mu_2} + \frac{d^2}{4}\right] \int_{R_z} (u_{3,t})^2 dx, \tag{5.7}$$

where $d =$ diameter D_0 and $h_0 = \min\{x_\alpha n_\alpha\}$ on ∂D_ξ.

Proof. We observe that from

$$\int_{R_z} x_\alpha\varphi_{,\alpha}\left[\Delta\varphi - u_{3,t}\right] dx = 0,$$

one obtains the Rellich-type identity

$$\frac{1}{2}\int_z^\infty \int_{\partial D_\xi} x_\alpha n_\alpha |\text{grad}_s\varphi|^2 ds d\xi = \int_{R_z} (\varphi_{,3})^2 dx - \int_{R_z} x_\alpha\varphi_{,\alpha}u_{3,t}dx$$

through integrating by parts twice and using the boundary condition on φ. By Schwarz's inequality, the AG inequality, and (5.5), we have

$$\frac{1}{2}\int_z^\infty \int_{\partial D_\xi} x_\alpha n_\alpha |\text{grad}_s\varphi|^2 ds d\xi \leq \left[\frac{1}{\mu_2} + \frac{d^2}{4}\right] \int_{R_z} (u_{3,t})^2 dx,$$

from which (5.7) follows.

Our next lemma which is needed to estimate the sixth integral in (5.3) does not depend on the auxiliary function φ, but rather on the solutions of the systems (2.1)–(2.6). We assert the following.

LEMMA 4. *For sufficiently smooth functions u_i and p in the system* (2.1)–(2.6), *the following identity holds*:

$$\begin{aligned}\frac{1}{2}\int_z^\infty \int_{\partial D_\xi} x_\alpha n_\alpha u_{i,j}\left(u_{i,j}-u_{j,i}\right) ds d\xi &= \int_{D_z} x_\alpha u_{\beta,\alpha}\left(u_{\beta,3}-u_{3,\beta}\right) dA \\ &-\frac{1}{2}\int_{R_z} u_{i,j}\left(u_{i,j}-u_{j,i}\right) dx + \int_{R_z} x_\alpha u_{i,\alpha} u_{i,t} dx \\ &+\int_{R_z}\left(x_3-z\right) u_{i,3} u_{i,t} dx + \int_{D_z} x_\alpha u_{3,\alpha} p dA.\end{aligned} \tag{5.8}$$

Proof. By (2.1), we know

$$\int_{R_z}\left(x_k - z\delta_{3k}\right) u_{i,k}\left[\left(u_{i,j}-u_{j,i}\right)_{,j} - u_{i,t} - p_{,i}\right] dx = 0,$$

where δ_{ij} is the Kronecker delta symbol. We integrate by parts in the first and third integrals and obtain

$$\begin{aligned}&\int_{\partial R_z}\left(x_k - z\delta_{3k}\right) u_{i,k}\left(u_{i,j}-u_{j,i}\right) n_j ds - \int_{R_z}\left(x_k - z\delta_{3k}\right) u_{i,jk}\left(u_{i,j}-u_{j,i}\right) dx \\ &-\int_{R_z} u_{i,j}\left(u_{i,j}-u_{j,i}\right) dx - \int_{R_z}\left(x_k - z\delta_{3k}\right) u_{i,k} u_{i,t} dx - \int_{\partial R_z}\left(x_k - z\delta_{3k}\right) u_{i,k} p n_i ds = 0.\end{aligned}$$

Now since

$$\left(u_{i,j}-u_{j,i}\right) u_{i,jk} = \left(u_{i,j}-u_{j,i}\right)_{,k} u_{i,j},$$

a further integration by parts results in

$$\begin{aligned}&\int_{\partial R_z}\left(x_k - z\delta_{3k}\right) u_{i,k}\left(u_{i,j}-u_{j,i}\right) n_j ds \\ &\quad-\frac{1}{2}\int_{\partial R_z}\left(x_k - z\delta_{3k}\right) u_{i,j}\left(u_{i,j}-u_{j,i}\right) n_k ds \\ &\quad+\frac{1}{2}\int_{R_z} u_{i,j}\left(u_{i,j}-u_{j,i}\right) dx - \int_{R_z}\left(x_k - z\delta_{3k}\right) u_{i,k} u_{i,t} dx \\ &\quad-\int_{\partial R_z}\left(x_k - z\delta_{3k}\right) u_{i,k} p n_i ds = 0.\end{aligned} \tag{5.9}$$

Since $u_{i,k}n_i = u_{i,i}n_k = 0$ on ∂D_ξ for $\xi \geq 0$, the last term in (5.9) can be written

$$\int_{\partial R_z}\left(x_k - z\delta_{3k}\right) u_{i,k} p n_i ds = -\int_{D_z} x_\alpha u_{3,\alpha} p dA.$$

Moreover, by (2.3), the first two integrals in (5.9) on the lateral boundary can be combined so that

$$\begin{aligned}&\int_z^\infty \int_{\partial D_\xi}\left(x_k - z\delta_{3k}\right) u_{i,k}\left(u_{i,j}-u_{j,i}\right) n_j ds d\xi \\ &\quad-\frac{1}{2}\int_z^\infty \int_{\partial D_\xi}\left(x_k - z\delta_{3k}\right) u_{i,j}\left(u_{i,j}-u_{j,i}\right) n_k ds d\xi \\ &=\frac{1}{2}\int_z^\infty \int_{\partial D_\xi} x_\alpha n_\alpha u_{i,j}\left(u_{i,j}-u_{j,i}\right) ds d\xi.\end{aligned} \tag{5.10}$$

Consequently, solving (5.9) for the condensed term obtained in (5.10), we arrive at the identity (5.8).

The appearance of the pressure function p in the identity (5.8) presents a further complication that must now be handled. We define $\overline{p}$ to be the mean value of p over D_z, i.e.,

$$\overline{p} = \frac{1}{|D_z|} \int_{D_z} p dA,$$

where $|D_z| = |D_0|$ is the measure of D_z. It follows that

$$\int_{D_z} [p - \overline{p}]\, dA = 0 \tag{5.11}$$

for each $z \geq 0$. Further, we note that the last term in (5.8) can be replaced by

$$\begin{aligned} \int_{D_z} x_\alpha u_{3,\alpha} p dA &= \int_{D_z} x_\alpha u_{3,\alpha} (p - \overline{p})\, dA + \int_{D_z} x_\alpha u_{3,\alpha} \overline{p} dA \\ &= \int_{D_z} x_\alpha u_{3,\alpha} (p - \overline{p})\, dA, \end{aligned} \tag{5.12}$$

since

$$\int_{D_z} x_\alpha u_{3,\alpha} dA = -2 \int_{D_z} u_3 dA = 0.$$

Since

$$u_{i,j}(u_{i,j} - u_{j,i}) = \tfrac{1}{2}(u_{i,j} - u_{j,i})(u_{i,j} - u_{j,i}),$$

from (5.8), (5.12), and Schwarz's inequality, we have

$$\begin{aligned} \frac{h_0}{2} \int_z^\infty \int_{\partial D_\xi} u_{i,j} (u_{i,j} - u_{j,i})\, ds d\xi &\leq \left(d \int_{D_z} u_{\beta,\alpha} u_{\beta,\alpha} dA \right)^{\frac{1}{2}} \\ &\quad \times \left(d \int_{D_z} (u_{\alpha,3} - u_{3,\alpha})\, (u_{\alpha,3} - u_{3,\alpha})\, dA \right)^{\frac{1}{2}} \\ &\quad + \left(d^2 \int_{R_z} u_{i,t} u_{i,t} dx \right)^{\frac{1}{2}} \left(\int_{R_z} u_{i,\alpha} u_{i,\alpha} dx \right)^{\frac{1}{2}} \\ &\quad + \left(\int_{R_z} (\xi - z)\, u_{i,3} u_{i,3} dx \right)^{\frac{1}{2}} \left(\int_{R_z} (\xi - z)\, u_{i,t} u_{i,t} dx \right)^{\frac{1}{2}} \\ &\quad + \left(d \int_{D_z} u_{3,\alpha} u_{3,\alpha} dA \right)^{\frac{1}{2}} \left(d \int_{D_z} (p - \overline{p})^2\, dA \right)^{\frac{1}{2}}. \end{aligned} \tag{5.13}$$

The need to bound the last integral in (5.13) motivates our second auxiliary problem.

We consider the boundary value problem

$$\begin{cases} \Delta \Psi = 0 & \text{in } R_z, \\ \dfrac{\partial \Psi}{\partial n} = 0 & \text{on } \partial D_\xi, \quad \xi \geq z \geq 0, \\ \dfrac{\partial \Psi}{\partial n} = p - \overline{p} & \text{in } D_z. \end{cases} \tag{5.14}$$

We note by (5.11) that the necessary condition for existence of a solution Ψ is satisfied and recall that Ψ is defined up to an arbitrary constant. We choose the constant such that

$$\lim_{z\to\infty}\int_{D_z}\Psi dA=0.$$

Then from the Phragmèn–Lindelöf theory, we know that there exists a solution Ψ which vanishes exponentially as $x_3\to\infty$. In addition, Ψ has the following properties.

LEMMA 5. *If Ψ is chosen as above, then for all $z\geq 0$,*

$$\int_{D_z}\Psi dA=0, \tag{5.15}$$

$$\int_{D_z}\Psi_{,3}dA=0, \tag{5.16}$$

$$\int_{R_z}\overline{p}_{,3}\Psi_{,3}dx=0. \tag{5.17}$$

Proof. The first property (5.15) follows by an argument similar to Lemma 1. For (5.16), we see that

$$\begin{aligned}\int_{D_z}\Psi_{,3}dA&=-\int_z^\infty\int_{D_\xi}\Psi_{,33}dAd\xi=\int_z^\infty\int_{D_\xi}\Psi_{,\alpha\alpha}dAd\xi\\&=\int_z^\infty\int_{\partial D_\xi}\Psi_{,\alpha}n_\alpha dsd\xi=0.\end{aligned}$$

The third property (5.17) follows from (5.16) since we can write

$$\int_{R_z}\overline{p}_{,3}\Psi_{,3}dx=\int_z^\infty\left(\frac{1}{|D_\xi|}\int_{D_\xi}p_{,3}dA\int_{D_\xi}\Psi_{,3}dA\right)d\xi.$$

By means of the auxiliary problem (5.14) and property (5.17), we have

$$\begin{aligned}\int_{D_z}(p-\overline{p})^2dA&=\int_{\partial R_z}(p-\overline{p})\frac{\partial\Psi}{\partial n}ds\\&=\int_{R_z}(p-\overline{p})_{,i}\Psi_{,i}dx\\&=\int_{R_z}\Psi_{,i}[(u_{i,j}-u_{j,i})_{,j}-u_{i,t}]dx.\end{aligned}$$

Then, in a manner similar to the derivation of (5.3), we can write

$$\begin{aligned}\int_{D_z}(p-\overline{p})^2dA\leq{}&\left(\int_{R_z}\Psi_{,i}\Psi_{,i}dx\right)^{\frac12}\left(\int_{R_z}u_{i,t}u_{i,t}dx\right)^{\frac12}\\&+\left(\int_{D_z}\Psi_{,\alpha}\Psi_{,\alpha}dA\right)^{\frac12}\left(\int_{D_z}(u_{\alpha,3}-u_{3,\alpha})(u_{\alpha,3}-u_{3,\alpha})dA\right)^{\frac12}\\&+\left(\int_z^\infty\int_{\partial D_\xi}|\mathrm{grad}_s\Psi|^2dsd\xi\right)^{\frac12}\left(\int_z^\infty\int_{\partial D_\xi}(u_{i,j}-u_{j,i})u_{i,j}dsd\xi\right)^{\frac12}.\end{aligned}\tag{5.18}$$

We now seek estimates for the integrals involving the function Ψ. These are established in the following lemmas.

LEMMA 6. *The function Ψ satisfies*

$$\int_{D_z} \Psi_{,\alpha}\Psi_{,\alpha} dA = \int_{D_z} (p-\bar{p})^2\, dA, \tag{5.19}$$

$$\int_{R_z} \Psi_{,i}\Psi_{,i} dx \le \frac{1}{\sqrt{\mu_2}} \int_{D_z} (p-\bar{p})^2\, dA. \tag{5.20}$$

Proof. For (5.19), we observe that

$$\int_{R_z} \Psi_{,3}\Delta\Psi dx = 0$$

implies the identity

$$\int_{D_z} (\Psi_{,3})^2 dA - \frac{1}{2}\int_{D_z} \Psi_{,j}\Psi_{,j} dA = 0,$$

and hence

$$\int_{D_z} \Psi_{,\alpha}\Psi_{,\alpha} dA = \int_{D_z} (\Psi_{,3})^2 dA = \int_{D_z} (p-\bar{p})^2\, dA.$$

The second result (5.20) follows from (5.14), (5.15), (5,19), and the calculation

$$\int_{R_z} \Psi_{,i}\Psi_{,i} dx = \int_{D_z} \Psi\,(p-\bar{p})\, dA \le \left(\frac{1}{\mu_2}\int_{D_z} \Psi_{,\alpha}\Psi_{,\alpha} dA\right)^{\frac{1}{2}} \left(\int_{D_z} (p-\bar{p})^2\, dA\right)^{\frac{1}{2}}.$$

LEMMA 7. *If D_z is star-shaped with respect to a point* (*origin*) *in D_z, then*

$$\int_z^{\infty}\int_{\partial D_\xi} |\text{grad}_s \Psi|^2\, ds d\xi \le \frac{2}{h_0}\left[d + \frac{1}{\sqrt{\mu_2}}\right]\int_{D_z} (p-\bar{p})^2\, dA. \tag{5.21}$$

Proof. In a manner similar to the proof of Lemma 3, from

$$\int_{R_z} x_\alpha \Psi_{,\alpha}\Delta\Psi dx = 0$$

we obtain the identity

$$\frac{1}{2}\int_z^{\infty}\int_{\partial D_\xi} x_\alpha n_\alpha\, |\text{grad}_s\Psi|^2\, ds d\xi = \int_{D_z} x_\alpha \Psi_{,\alpha}\,(p-\bar{p})\, dA + \int_{R_z} (\Psi_{,3})^2 dx,$$

and hence the inequality

$$\frac{h_0}{2}\int_z^{\infty}\int_{\partial D_\xi} |\text{grad}_s\Psi|^2\, ds d\xi \le \frac{d}{2}\int_{D_z} \Psi_{,\alpha}\Psi_{,\alpha} dA + \frac{d}{2}\int_{D_z} (p-\bar{p})^2\, dA + \int_{R_z} \Psi_{,i}\Psi_{,i} dx.$$

Now by (5.19) and (5.20), the result follows.

We can use Lemmas 6 and 7 in (5.18) to obtain an estimate for the integral of

$(p-\bar{p})^2$ which is independent of pressure, namely,

$$\begin{aligned}\left(\int_{D_z}(p-\bar{p})^2 dA\right)^{\frac{1}{2}} &\leq \left(\frac{1}{\sqrt{\mu_2}}\int_{R_z} u_{i,t}u_{i,t}dx\right)^{\frac{1}{2}}\\ &+\left(\int_{D_z}(u_{\alpha,3}-u_{3,\alpha})(u_{\alpha,3}-u_{3,\alpha})dA\right)^{\frac{1}{2}}\\ &+\left(\frac{2}{h_0}\left[d+\frac{1}{\sqrt{\mu_2}}\right]\int_z^\infty\int_{\partial D_\xi}(u_{i,j}-u_{j,i})u_{i,j}dsd\xi\right)^{\frac{1}{2}}.\end{aligned} \tag{5.22}$$

We are now ready to complete an estimate for the I_4 integral.

We return to inequality (5.3) and use the weighted AG inequality, Lemma 2, and Lemma 3 to obtain

$$\begin{aligned}I_4 \leq &\frac{\gamma_1}{2\mu_2}\int_0^t\int_{R_z} k(u_{3,\tau})^2 dxd\tau+\frac{1}{2\gamma_1}\int_0^t\int_{R_z} ku_{i,\tau}u_{i,\tau}dxd\tau\\ &+\frac{\gamma_2}{\sqrt{\mu_2}}\int_0^t\int_{R_z} k(u_{3,\tau})^2 dxd\tau+\frac{k}{2\gamma_2}\int_0^t\int_{D_z}(u_{\alpha,3}-u_{3,\alpha})(u_{\alpha,3}-u_{3,\alpha})dAd\tau\\ &+\frac{k\gamma_3}{h_0}\left[\frac{1}{\mu_2}+\frac{d^2}{4}\right]\int_0^t\int_{R_z} k(u_{3,\tau})^2 dxd\tau\\ &+\frac{1}{2\gamma_3}\int_0^t\int_z^\infty\int_{\partial D_\xi}(u_{i,j}-u_{j,i})u_{i,j}dsd\xi d\tau.\end{aligned}$$

Since $k=1/\lambda_1$, we choose

$$\gamma_1=\sqrt{\mu_2},\qquad \gamma_2=1/2\,,\qquad \gamma_3=\lambda_1/\gamma,$$

where γ is a dimensionless constant to be determined, so that

$$\begin{aligned}I_4 \leq &\left(\frac{3}{2\sqrt{\mu_2}}+\frac{1}{\gamma h_0}\left[\frac{1}{\mu_2}+\frac{d^2}{4}\right]\right)\int_0^t\int_{R_z} ku_{i,\tau}u_{i,\tau}dxd\tau\\ &+\frac{1}{\lambda_1}\int_0^t\int_{D_z}(u_{\alpha,3}-u_{3,\alpha})(u_{\alpha,3}-u_{3,\alpha})dAd\tau+\frac{\gamma}{2\lambda_1}J,\end{aligned} \tag{5.23}$$

where

$$J=\int_0^t\int_z^\infty\int_{\partial D_\xi}(u_{i,j}-u_{j,i})u_{i,j}dsd\xi d\tau.$$

The first term in (5.23) is estimable in terms of $-\partial E/\partial z$. For the second term, we write

$$\begin{aligned}&\frac{1}{\lambda_1}\int_0^t\int_{D_z}(u_{\alpha,3}-u_{3,\alpha})(u_{\alpha,3}-u_{3,\alpha})dAd\tau\\ &\leq \frac{1}{\lambda_1}(1+\eta)\int_0^t\int_{D_z} u_{\alpha,3}u_{\alpha,3}dAd\tau\\ &+\frac{1}{\lambda_1}\left(1+\frac{1}{\eta}\right)\int_0^t\int_{D_z} u_{3,\alpha}u_{3,\alpha}dAd\tau,\end{aligned} \tag{5.24}$$

where η is a constant to be suitably chosen. To estimate the third term in (5.23), we recall (5.13) and use the weighted AG inequality together with (5.22) on the last term of (5.13) to obtain

$$\begin{aligned}\frac{h_0}{4}\int_z^\infty \int_{\partial D_\xi} (u_{i,j}-u_{j,i})\,u_{i,j}\,ds\,d\xi \le{}& \frac{d\gamma_4}{2}\int_{D_z} u_{\beta,\alpha}u_{\beta,\alpha}dA \\ &+ \frac{d}{2\gamma_4}\int_{D_z}(u_{\alpha,3}-u_{3,\alpha})\,(u_{\alpha,3}-u_{3,\alpha})\,dA \\ &+ \frac{d^2\gamma_5}{2}\int_{R_z} u_{i,t}u_{i,t}dx + \frac{1}{2\gamma_5}\int_{R_z}u_{i,\alpha}u_{i,\alpha}dx \\ &+ \frac{\gamma_6}{2}\int_{R_z}(\xi-z)\,u_{i,3}u_{i,3}dx + \frac{1}{2\gamma_6}\int_{R_z}(\xi-z)\,u_{i,t}u_{i,t}dx \\ &+ \frac{d}{2}(\gamma_7+\gamma_8+\epsilon)\int_{D_z}u_{3,\alpha}u_{3,\alpha}dA + \frac{d}{2\gamma_7\sqrt{\mu_2}}\int_{R_z}u_{i,t}u_{i,t}dx \\ &+ \frac{d}{2\gamma_8}\int_{D_z}(u_{\alpha,3}-u_{3,\alpha})\,(u_{\alpha,3}-u_{3,\alpha})\,dA,\end{aligned}\tag{5.25}$$

where we have set

$$\epsilon = \frac{4d}{h_0^2}\left[d+\frac{1}{\sqrt{\mu_2}}\right].$$

We now integrate (5.25) with respect to τ, multiply by $2\gamma/\lambda_1 h_0$, and then add the resulting inequality to (5.24). Collecting terms, we have

(5.26)

$$\begin{aligned}&\frac{1}{\lambda_1}\int_0^t\int_{D_z}(u_{\alpha,3}-u_{3,\alpha})\,(u_{\alpha,3}-u_{3,\alpha})\,dA\,d\tau + \frac{\gamma}{2\lambda_1}J \\ &\le \frac{1}{\lambda_1}(1+\eta)\left(1+\frac{\gamma d}{h_0}\left[\frac{1}{\gamma_4}+\frac{1}{\gamma_8}\right]\right)\int_0^t\int_{D_z}u_{\alpha,3}u_{\alpha,3}dA\,d\tau \\ &\quad+\frac{1}{\lambda_1}\left(\left[1+\frac{1}{\eta}\right]\left[1+\frac{\gamma d}{h_0}\left(\frac{1}{\gamma_4}+\frac{1}{\gamma_8}\right)\right]+\frac{\gamma d}{h_0}\left[\gamma_7+\gamma_8+\epsilon\right]\right)\int_0^t\int_{D_z}u_{3,\alpha}u_{3,\alpha}dA\,d\tau \\ &\quad+\frac{\gamma d\gamma_4}{\lambda_1 h_0}\int_0^t\int_{D_z}u_{\beta,\alpha}u_{\beta,\alpha}dA\,d\tau + \frac{\gamma d}{\lambda_1 h_0\gamma_7\sqrt{\mu_2}}\int_0^t\int_{R_z}u_{i,\tau}u_{i,\tau}dx\,d\tau \\ &\quad+\frac{\gamma}{\lambda_1 h_0}\left\{\gamma_5 d^2\int_0^t\int_{R_z}u_{i,\tau}u_{i,\tau}dx\,d\tau + \frac{1}{\gamma_5}\int_0^t\int_{R_z}u_{i,\alpha}u_{i,\alpha}dx\,d\tau\right\} \\ &\quad+\frac{\gamma}{\lambda_1 h_0}\left\{\gamma_6\int_0^t\int_{R_z}(\xi-z)\,u_{i,3}u_{i,3}dx\,d\tau + \frac{1}{\gamma_6}\int_0^t\int_{R_z}(\xi-z)\,u_{i,\tau}u_{i,\tau}dx\,d\tau\right\}.\end{aligned}$$

We recall $k = 1/\lambda_1$ and choose

$$\gamma_4=\gamma_8=\frac{d\sqrt{\lambda_1}}{2},\qquad \gamma_5=\frac{1}{d\sqrt{2\lambda_1}},\qquad \gamma_6=\sqrt{\lambda_1},\qquad \gamma_7=\sqrt{\frac{2\lambda_1}{\mu_2}},\qquad \gamma=\frac{h_0\sqrt{\lambda_1}}{2}.$$

Moreover, we choose η to be the positive root of the quadratic equation

$$\eta^2 - c\eta - 1 = 0,\tag{5.27}$$

where

$$c=\frac{d\sqrt{\lambda_1}}{6}\left[\sqrt{\frac{2\lambda_1}{\mu_2}}+\frac{d\sqrt{\lambda_1}}{2}+\frac{4d^2}{h_0^2}+\frac{4d}{h_0^2\sqrt{\mu_2}}\right],$$

so that the coefficients of the first two integrals on the right side of (5.26) are equal. Thus, we can write

(5.28)

$$\frac{1}{\lambda_1}\int_0^t\int_{D_z}\left(u_{\alpha,3}-u_{3,\alpha}\right)\left(u_{\alpha,3}-u_{3,\alpha}\right)dAd\tau+\frac{\gamma}{2\lambda_1}J\le\rho\frac{\partial^2 E}{\partial z^2}+\frac{d}{\sqrt{2}}\left(-\frac{\partial E}{\partial z}\right)+\frac{1}{2}E,$$

where

$$\rho=\max\left\{\frac{d^2}{4},\frac{3}{\lambda_1}(1+\eta)\right\}. \tag{5.29}$$

We combine inequalities (5.23) and (5.28) to obtain our estimate on I_4, namely,

$$I_4\le\left(\frac{3}{2\sqrt{\mu_2}}+\frac{2}{h_0^2\sqrt{\lambda_1}}\left[\frac{1}{\mu_2}+\frac{d^2}{4}\right]+\frac{d\sqrt{2}}{2}\right)\left(-\frac{\partial E}{\partial z}\right)+\rho\frac{\partial^2 E}{\partial z^2}+\frac{1}{2}E. \tag{5.30}$$

Consequently, by combining (4.2), (4.8), and (5.30), we obtain the second-order differential inequality

$$E\le K_1\frac{\partial^2 E}{\partial z^2}-K_2\frac{\partial E}{\partial z}, \tag{5.31}$$

where

$$K_1=\frac{\sqrt{C}}{\lambda_1}+2\rho,$$

$$K_2=2\left(\frac{1+\sqrt{C}}{\sqrt{\lambda_1}}\right)+\frac{3}{\sqrt{\mu_2}}+\frac{4}{\mu_2 h_0^2\sqrt{\lambda_1}}+\frac{d^2}{h_0^2\sqrt{\lambda_1}}+d\sqrt{2}.$$

We can rewrite (5.31) as

$$\frac{\partial^2 E}{\partial z^2}-a\frac{\partial E}{\partial z}-bE\ge 0, \tag{5.32}$$

where $a=K_2/K_1$ and $b=1/K_1$. Furthermore, we can write (5.32) in the form

$$\left(\frac{\partial}{\partial z}-k_1\right)\left(\frac{\partial E}{\partial z}+k_2E\right)\ge 0, \tag{5.33}$$

where

$$k_1=\frac{a}{2}+\frac{1}{2}\sqrt{a^2+4b},\qquad k_2=-\frac{a}{2}+\frac{1}{2}\sqrt{a^2+4b}.$$

Now since

$$\frac{\partial}{\partial z}\left\{e^{-k_1 z}\left(\frac{\partial E}{\partial z}+k_2E\right)\right\}\ge 0,$$

we conclude upon integration from z to ∞ that

$$\frac{\partial E}{\partial z}+k_2E\le 0,$$

and hence that

$$E(z,t)\le E(0,t)\,e^{-k_2 z}. \tag{5.34}$$

This is the exponential decay estimate we set out to obtain.

We note that k_2 depends only on geometry and the eigenvalues λ_1 and μ_2. In fact, using $\lambda_1 = 5.78/r^2$ and $\mu_2 = 3.38/r^2$ for a circular pipe of radius r, we find $k_2 \approx 0.15/r$. We note further that the decay constant k_2 might be improved by a more judicious choice of weights in our derivation above. To complete the energy estimation, we will indicate a procedure for obtaining estimates on the total weighted energy $E\,(0,t)$ in the next section.

6. Bound for $E(0,t)$. We now indicate how one can estimate the total weighted energy. We shall not determine explicit constants in this estimation as we did in the previous sections. Here we let ϵ_i designate positive constants that may be chosen arbitrarily small and let c_i denote computable constants that may depend on the ϵ_i but which remain bounded if any of the ϵ_i tend to zero.

We first note from (5.31),

$$E\,(0,t) \leq K_1 \frac{\partial^2 E}{\partial z^2}\,(0,t) - K_2 \frac{\partial E}{\partial z}\,(0,t)\,. \tag{6.1}$$

Moreover, from (2.9),

$$\begin{aligned} \frac{\partial^2 E}{\partial z^2}\,(0,t) &= \text{data } + \int_0^t \int_{D_0} u_{i,3} u_{i,3} dA d\tau \\ &= \text{data } + \int_0^t \int_{D_0} u_{\alpha,3} u_{\alpha,3} dA d\tau + \int_0^t \int_{D_0} \left(f_{\alpha,\alpha}\right)^2 dA d\tau \end{aligned} \tag{6.2}$$

since $u_{3,3} = -u_{\alpha,\alpha} = -f_{\alpha,\alpha}$ on D_0. Thus we must bound

$$-\frac{\partial E}{\partial z}\,(0,t)\,, \qquad \int_0^t \int_{D_0} u_{\alpha,3} u_{\alpha,3} dA d\tau \tag{6.3}$$

suitably to complete the estimation for $E\,(0,t)$.

Consider the first term in (6.3). On integration by parts, we have

$$\begin{aligned} \int_0^t \int_{R_0} u_{i,j} u_{i,j} dx d\tau &\leq - \int_0^t \int_{D_0} f_i u_{i,3} dA d\tau + \int_0^t \int_{D_0} u_3 p dA d\tau \\ &= - \int_o^t \int_{D_0} f_\alpha u_{\alpha,3} dA d\tau + \int_0^t \int_{D_0} f_3 f_{\alpha,\alpha} dA d\tau \\ &\quad + \int_0^t \int_{D_0} f_3\,(p - \overline{p})\, dA d\tau, \end{aligned} \tag{6.4}$$

and that

$$\begin{aligned} \int_0^t \int_{R_0} u_{i,\tau} u_{i,\tau} dx d\tau &= \int_0^t \int_{R_0} \left[\Delta u_i - p_{,i}\right] u_{i,\tau} dx d\tau \\ &\leq - \int_0^t \int_{D_0} \left[f_{i,\tau} u_{i,3} - u_{3,\tau}\,(p - \overline{p})\right] dA d\tau \\ &= - \int_0^t \int_{D_0} f_{\alpha,\tau} u_{\alpha,3} dA d\tau + \int_0^t \int_{D_0} f_{3,\tau} f_{\alpha,\alpha} dA d\tau \\ &\quad + \int_0^t \int_{D_0} f_{3,\tau}\,(p - \overline{p})\, dA d\tau. \end{aligned} \tag{6.5}$$

Thus, combining (6.4) and (6.5), we find

$$-\frac{\partial E}{\partial z}(0,t) \leq \text{data } + \epsilon_1 \int_0^t \int_{D_0} u_{\alpha,3}u_{\alpha,3}dAd\tau + \epsilon_2 \int_0^t \int_{D_0} (p-\overline{p})^2 \, dAd\tau, \tag{6.6}$$

where ϵ_1 and ϵ_1 are constants to be chosen appropriately.

We now observe that from (5.22) and (5.26), we can write

$$\begin{aligned}\int_0^t \int_{D_0} (p-\overline{p})^2 \, dAd\tau \leq & \text{data} + \; c_1 \int_0^t \int_{D_0} u_{\alpha,3}u_{\alpha,3}dAd\tau \\ & - c_2 \frac{\partial E}{\partial z}(0,t) + c_3 E(0,t).\end{aligned} \tag{6.7}$$

By means of (6.6) and an appropriate (first restriction on) choice of ϵ_2, we have

$$\int_0^t \int_{D_0} (p-\overline{p})^2 \, dAd\tau \leq \text{data } + c_4 \int_0^t \int_{D_0} u_{\alpha,3}u_{\alpha,3}dAd\tau + c_5 E(0,t). \tag{6.8}$$

Substituting (6.8) into (6.6), we obtain

$$-\frac{\partial E}{\partial z}(0,t) \leq \text{data } + (\epsilon_1 + \epsilon_2 c_4) \int_0^t \int_{D_0} u_{\alpha,3}u_{\alpha,3}dAd\tau + \epsilon_2 c_5 E(0,t). \tag{6.9}$$

Thus, combining (6.1), (6.2), and (6.9), we can write

$$E(0,t) \leq \text{data } + c_6 \int_0^t \int_{D_0} u_{\alpha,3}u_{\alpha,3}dAd\tau + \epsilon_2 c_7 E(0,t), \tag{6.10}$$

and it only remains to bound the middle term on the right side of (6.10).

To establish a bound for the second term of (6.3), we consider the identity

$$\int_0^t \int_{R_0} u_{i,3} \left[(u_{i,j} - u_{j,i})_{,j} - p_{,i} - u_{i,\tau} \right] dxd\tau = 0.$$

On integrating by parts, we are led to

$$\begin{aligned}- \int_0^t \int_{D_0} u_{i,3}(u_{i,3} - u_{3,i}) \, dAd\tau + \frac{1}{2} \int_0^t \int_{D_0} u_{i,j}(u_{i,j} - u_{j,i}) \, dAd\tau \\ + \int_0^t \int_{D_0} u_{3,3} p dAd\tau - \int_0^t \int_{R_0} u_{i,\tau} u_{i,3} dxd\tau = 0.\end{aligned}$$

Since $u_{3,3} = -f_{\alpha,\alpha}$ on D_0, we can now write

$$\begin{aligned}\int_0^t \int_{D_0} \left[u_{\alpha,3}(u_{\alpha,3} - u_{3,\alpha}) - \frac{1}{2} u_{\alpha,3}(u_{\alpha,3} - u_{3,\alpha}) - \frac{1}{2} f_{\alpha,\beta}(f_{\alpha,\beta} - f_{\beta,\alpha}) \right] dAd\tau \\ \leq \int_0^t \int_{D_0} f_{\alpha,\alpha}(p - \overline{p}) \, dAd\tau - c_8 \frac{\partial E}{\partial z}(0,t),\end{aligned}$$

or further,

$$\begin{aligned}\frac{1}{2} \int_0^t \int_{D_0} u_{\alpha,3}u_{\alpha,3}dAd\tau \leq \frac{1}{2} \int_0^t \int_{D_0} [f_{3,\alpha}f_{3,\alpha} + f_{\alpha,\beta}(f_{\alpha,\beta} - f_{\beta,\alpha})] \, dAd\tau \\ + \left(\int_0^t \int_{D_0} (f_{\alpha,\alpha})^2 dAd\tau \int_0^t \int_{D_0} (p-\overline{p})^2 \, dAd\tau \right)^{\frac{1}{2}} - c_8 \frac{\partial E}{\partial z}(0,t).\end{aligned}$$

It follows from (6.8) and (6.9) that

$$\int_0^t \int_{D_0} u_{\alpha,3}u_{\alpha,3}dAd\tau$$
$$\leq \text{data } + (\epsilon_3 c_4 + c_8 [\epsilon_1 + \epsilon_2 c_4]) \int_0^t \int_{D_0} u_{\alpha,3}u_{\alpha,3}dAd\tau + (\epsilon_3 c_5 + \epsilon_2 c_9) E(0,t).$$

With an appropriate choice for ϵ_1, ϵ_2, and ϵ_3 (a second restriction on ϵ_2), we have

$$\int_0^t \int_{D_0} u_{\alpha,3}u_{\alpha,3}dAd\tau \leq \text{data } + (\epsilon_3 c_{10} + \epsilon_2 c_{11}) E(0,t). \tag{6.11}$$

We complete our estimation of the total weighted energy by combining (6.10) and (6.11) and putting further restrictions on ϵ_2 and ϵ_3. Thus it follows that

$$E(0,t) \leq \text{data}, \tag{6.12}$$

as we set out to show.

7. Concluding remarks. Instead of considering the slow flow of an incompressible viscous fluid in a semi-infinite cylinder as we did here, one might consider the finite pipe problem. In this case, if one assumes

$$u_i = 0 \quad \text{at } x_3 = \ell,$$

for the pipe of length ℓ, then the analysis of the preceding sections applies with appropriate modifications.

One might also weaken the hypotheses which we imposed in order to obtain sufficient decay at infinity for the semi-infinite pipe. It appears that

$$u_i = o(1), \quad u_{i,t} = o(1), \quad [u_{\alpha,3}u_{\alpha,3}]^{\frac{1}{2}} = o\left(x_3^{-1}\right), \quad p = o\left(x_3^{-1}\right),$$

uniformly in x_1, x_2, t as $x_3 \to \infty$ would be sufficient for the needed convergence.

Finally, the second-order differential inequality (5.31) could be handled differently. Instead of (5.33), suppose we write

$$\left(\frac{\partial}{\partial z} + k_2\right)\left(\frac{\partial E}{\partial z} - k_1 E\right) \geq 0. \tag{7.1}$$

Then

$$\frac{\partial}{\partial z} e^{k_2 z}\left(\frac{\partial E}{\partial z} - k_1 E\right) \geq 0,$$

and integrating from 0 to z, we have

$$\frac{\partial E}{\partial z}(z,t) - k_1 E(z,t) \geq e^{-k_2 z}\left[\frac{\partial E}{\partial z}(0,t) - k_1 E(0,t)\right].$$

Consequently,

$$-\frac{\partial E}{\partial z}(z,t) + k_1 E(z,t) \leq e^{-k_2 z}\left[-\frac{\partial E}{\partial z}(0,t) + k_1 E(0,t)\right], \tag{7.2}$$

and we have established a bound for both $E(z,t)$ and $-\partial E(z,t)/\partial z$. Bounds for the terms on the right side of (7.2) can then be established as in §6.

Acknowledgment. The authors are grateful to a referee who suggested a simpler proof of Lemma 1.

REFERENCES

[1] K. A. AMES AND L. E. PAYNE, *Decay estimates in steady pipe flow*, SIAM J. Math Anal., 20 (1989), pp. 789–815.

[2] I. BABUŠKA AND A. K. AZIZ, *Survey lectures on the mathematical foundations of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, Academic Press, New York, 1972, pp. 5–359.

[3] B. A. BOLEY, *Some observations on Saint-Venant's principle*, in Proc. 3rd U.S. Nat. Cong. Appl. Mech., ASME, New York, 1958, pp. 259–264.

[4] A. R. ELCRAT AND V. G. SIGILLITO, *A spatial decay estimate for the Navier–Stokes equations*, J. Appl. Math. Phys. (ZAMP), 30 (1979), pp. 449–455.

[5] C. O. HORGAN AND J. K. KNOWLES, *Recent developments concerning Saint-Venant's principle*, Adv. Appl. Mech., 23 (1983), pp. 179–269.

[6] C. O. HORGAN AND L. E. PAYNE, *On inequalities of Korn, Friedrichs and Babuška–Aziz*, Arch. Rational Mech. Anal., 82 (1983), pp. 165–179.

[7] C. O. HORGAN AND L. T. WHEELER, *Spatial decay estimates for the Navier–Stokes equations with application to the problem of entry flow*, SIAM J. Appl. Math., 35 (1978), pp. 97–116.

[8] C. O. HORGAN, *Recent developments concerning Saint-Venant's principle: An update*, Appl. Mech. Rev, 42 (1989), pp. 295–303.

[9] C. LIN, *Spatial decay estimates and energy bounds for the Stokes flow equation*, preprint.

SIAM J. MATH. ANAL.
Vol. 24, No. 6, pp. 1414–1439, November 1993

A DEGENERATE PSEUDOPARABOLIC REGULARIZATION OF A NONLINEAR FORWARD-BACKWARD HEAT EQUATION ARISING IN THE THEORY OF HEAT AND MASS EXCHANGE IN STABLY STRATIFIED TURBULENT SHEAR FLOW*

G.I. BARENBLATT[†], M. BERTSCH[‡],
R. DAL PASSO[‡], AND M. UGHI[§]

Abstract. The authors analyze an initial-boundary value problem for the equation

$$u_t = \varphi(u_x)_x + \tau\psi(u_x)_{xt},$$

where τ is a positive parameter, $\varphi, \psi : \mathbf{R} \to \mathbf{R}$, φ is nonmonotone, ψ is strictly increasing and uniformly bounded in $\mathbf{R}$, and $|\varphi'(p)| = O(\psi'(p))$ as $p \to \pm\infty$. The equation arises as a (new) model for turbulent heat or mass transfer in stably stratified shear flows, in which case u_x is nonnegative, $\varphi(p) > 0$ for $p > 0$, and $\varphi(0) = \varphi(+\infty) = 0$. Well-posedness is proved and, in the model case, the qualitative behavior of solutions is studied. In particular it is shown that smooth solutions may become discontinuous in finite time, and that such solutions converge to a piecewise constant spatial profile as $t \to \infty$. This behavior is in agreement with experimental observations and numerical computations.

Key words. forward-backward parabolic equation, degenerate pseudoparabolic equation, existence, uniqueness, discontinuous solutions, large-time behavior, turbulent heat and mass diffusion, stratification, shear flow, relaxation time

AMS subject classifications. 35K70, 35K65, 76F10, 80A20

1. Introduction.

One-dimensional diffusion processes are modeled by the conservation law

$$u_t + q_x = 0, \qquad q = -ku_x, \tag{1.1}$$

where q is the heat or mass flux, x is a spatial coordinate belonging to some real interval, $t > 0$ indicates time, subscripts denote partial differentiations, and k is, by definition, the diffusivity. If k is a given function k_0 of $u(x,t)$, $u_x(x,t)$, x and t, its nonnegativity does not imply the parabolicity of (1.1), since the product $u_x k_0(u, u_x, x, t)$ may be nonmonotone with respect to u_x.

In this paper, we study a mathematical model for heat or mass transfer in a stably stratified turbulent shear flow, where the temperature, respectively concentration u, satisfies an equation of the type (1.1) (we refer to §2 for a brief discussion of the model). Under fixed external conditions the *steady* diffusivity, i.e., the diffusivity in mechanical and thermal equilibrium, depends only on the value of the gradient:

$$k(x,t) = k_0(u_x(x,t)). \tag{1.2}$$

*Received by the editors December 11, 1992; accepted for publication (in revised form) March 1, 1993. This research was supported in part with funds provided by the National Science Foundation.

[†] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Silver Street, Cambridge, England CB3 9EW.

[‡] Dipartimento di Matematica, Università di Roma Tor Vergata, Via della Ricerca Scientifica, 00133 Roma, Italy.

[§] Istituto delle Scienze delle Costruzioni, Università di Trieste, Piazzale Europa 1, 34127 Trieste, Italy.

For the stably stratified turbulent shear flow, the effective temperature or mass diffusivity k_0 decreases very quickly at large values of the temperature gradient, respectively concentration gradient, and a typical choice for k_0 is

$$k_0(p) = \frac{A}{B+p^2}, \qquad (A, B > 0), \tag{1.3}$$

and therefore the product

$$\varphi(p) = pk_0(p) \tag{1.4}$$

is nonmonotone: $\varphi(p)$ is increasing for $0 < p < \alpha$ for some critical constant $\alpha > 0$ (in the model u_x is always assumed to be nonnegative), but $\varphi(p)$ is decreasing for $p > \alpha$. More generally, we consider smooth functions φ which satisfy, for some $\alpha > 0$,

$$\begin{cases} \varphi(0) = \varphi(+\infty) = 0 \\ 0 < \varphi(p) \leq \varphi(\alpha) \quad \text{for } p > 0 \end{cases} \tag{1.5}$$

(see Fig. 1).

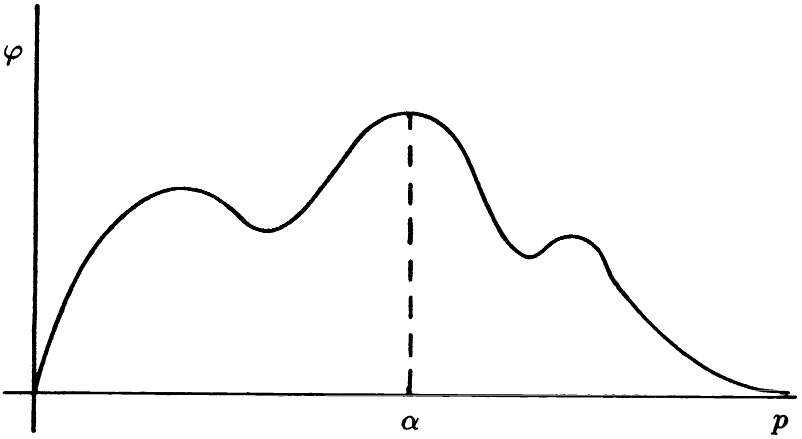


FIG. 1. *The function $\varphi(p)$ for $p \geq 0$.*

If we simply substitute (1.2) into the balance law (1.1), i.e., if we suppose that the turbulence field governing the thermal diffusivity is an equilibrium one, the resulting second-order partial differential equation

$$u_t = \varphi(u_x)_x \tag{1.6}$$

is not of forward parabolic type at points at which $\varphi'(u_x) < 0$.

If φ is not monotone, (1.6) leads to initial-boundary value problems which may be ill-posed. The ill-posedness does not necessarily regard the existence of solutions; for example, Höllig [H] showed that if φ is piecewise linear, decreasing in a bounded interval, and increasing elsewhere, there exist initial functions for which the corresponding Cauchy problem possesses infinitely many solutions.

A natural approach to treat a problem where an equation of the type (1.6) arises is to introduce some regularization which leads to well-posed problems. For example,

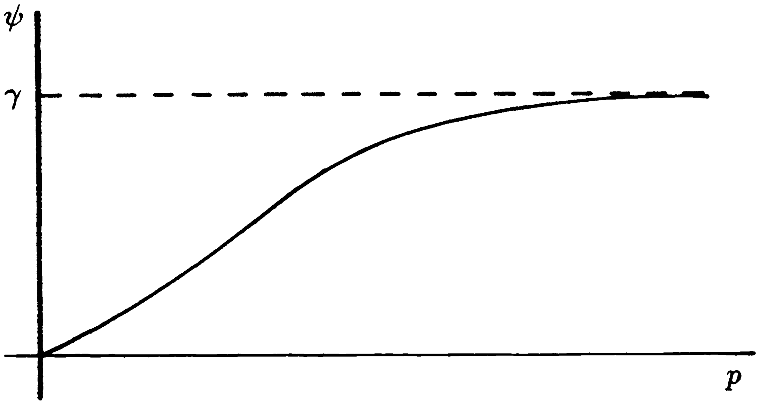


FIG. 2. *The function $\psi(p)$ for $p \geq 0$.*

one might add to the right-hand side of (1.6) the term $\epsilon^2 u_{xxt}$, which yields a pseudoparabolic equation, or $-\epsilon^2 u_{xxxx}$, which leads to a fourth-order parabolic equation. However, an important consequence of Höllig's nonuniqueness result is that the dynamics of the solutions (which will, of course, strongly depend on the type of nonlinear function φ under consideration) may critically depend on the sort of regularization which we introduce. Indeed, in the case of a cubic nonlinearity φ (i.e., φ is decreasing in a bounded interval, increasing elsewhere, and $\varphi(+\infty) = +\infty$), which arises in the context of viscoelasticity, the two regularizations which we mentioned before lead to different dynamics of solutions (we refer to the discussion in [NP]). This strongly suggests that, more than ever, an eventual regularization should be justified at the level of mathematical modeling. In other words, the regularization should be specific for the given physical problem. We observe that also in numerical computations a regularization appears, because difference equations are solved instead of differential equations (cf. the numerical computations in [Po], [Dj] for our model), and also in this context the consistency of the regularization with the physical model should be investigated.

In our model the natural way to modify the problem is to reconsider the assumption that the turbulence, which governs the diffusion coefficient and the flux, is in equilibrium with the temperature gradient. Indeed, in §2 we shall explain that if we take into account a small but positive time τ of the relaxation of the turbulence field to the temperature gradient, we obtain the equation

$$u_t = \varphi(u_x)_x + \tau\psi(u_x)_{xt}, \tag{1.7}$$

where ψ is a smooth nonlinear function which is expressed through φ (see Fig. 2). The properties of ψ which are important for us are

$$\begin{cases} \psi'(p) > 0 & \text{for } p > 0, \\ \psi(0) = 0, & \psi(+\infty) = \gamma < +\infty. \end{cases} \tag{1.8}$$

Since $\psi'(+\infty) = 0$, $\psi'(p)$ is not uniformly bounded away from zero, and we call (1.7) of *degenerate pseudoparabolic* type. In addition φ and ψ satisfy the inequality

$$\varphi'(p) \geq -\psi'(p) \quad \text{for } p > 0, \tag{1.9}$$

which will play an important role below since it implies, roughly speaking, that the last term in (1.7) is strong enough to control the possibly negative diffusion coefficient $\varphi'(u_x)$. For the precise hypotheses on the data we refer to §3.

Our main results concern the initial-boundary value problem

$$\text{(I)}\begin{cases} u_t = \varphi(u_x)_x + \tau\psi(u_x)_{xt} & \text{in } (0,1)\times\mathbf{R}^+, \\ u_x(0,t) = u_x(1,t) = 0 & \text{for } t > 0, \\ u(x,0) = u_0(x) & \text{for } 0 < x < 1. \end{cases}$$

The lateral boundary conditions in problem (I) are no-flux conditions, but the mathematical analysis which we shall present does not depend strongly on the choice of the boundary conditions; in particular, a similar analysis is possible for the corresponding Cauchy problem in which $-\infty < x < +\infty$.

The main results of this paper regard the well-posedness of problem (I), and both the transient and the large time behavior of its solutions. Below we shall briefly describe these results; for their precise statement we refer again to §3.

The uniqueness and the existence of a (generalized) solution will be proved in, respectively, §§4 and 5.

The qualitative transient behavior of the solutions depends strongly on the properties of the function φ and the threshold α for the gradient u_x. Let us assume for the moment that the initial function u_0 is smooth and nondecreasing in the interval $(0,1)$. In §6 we shall prove that $u_x \geq 0$ in $(0,1)\times\mathbf{R}^+$ and that, given $x_0 \in (0,1)$,

$$u_0'(x_0) \leq \alpha \Rightarrow u_x(x_0,t) \leq \alpha \quad \text{for all } t > 0.$$

In particular, if $u_0' \leq \alpha$ in $(0,1)$, then the gradient u_x is uniformly bounded in $(0,1)\times\mathbf{R}^+$.

The situation changes drastically if u_0' is not bounded by α. In §8 we shall show that there exists a smooth initial function such that the solution becomes *discontinuous within finite time.* Of course this phenomenon is caused by the specific form of φ, but we observe that, due the the strong degeneracy of $\psi(p)$ as $p \to \infty$, the regularizing effect of the third-order term is not sufficient to prevent the formation of discontinuities. In §7 we shall prove that, once a solution is discontinuous at some point x_0, it remains discontinuous at x_0 for all later times t; more precisely we shall show that the jump $u(x_0^+,t)-u(x_0^-,t)$ is nondecreasing with respect to t (actually the growth of the jump is the essential part of the result; the mere preservation of the discontinuity could be guessed from earlier results [NP], [S] on pseudoparabolic equations).

In §9 we shall prove that, given u_0, there exists a function $q \in BV((0,1))$ such that $u(\cdot,t) \to q$ almost everywhere in $(0,1)$ as $t \to \infty$, and the regular part of the derivative q' vanishes almost everywhere in $(0,1)$; the singular part of the measure q' does not necessarily vanish: if u is discontinuous at some point (x_0,t_0), q will also be discontinuous at x_0.

Physically such piecewise constant asymptotic profiles correspond to the formation of stepwise temperature or salinity distributions in, for example, the ocean. On the modeling level it was known for a long time that the nonmonotonicity of φ could explain the formation of steps as $t \to \infty$, but it was always tacitly assumed that $\tau = 0$, without paying attention to the mathematical well-posedness of the problem. As a conclusion we may say that our assumption of positive relaxation time τ guarantees well-posedness of the model and preserves the required qualitative behavior of solutions.

Finally we discuss briefly some of the literature about degenerate and nondegenerate pseudoparabolic equations. For a first introduction to the subject and a list of references to early papers we refer to Chapter 3 in Carroll and Showalter [CS]. Among the more recent papers we mention [BG], [BS], [dBP], [N], [NP], [P], [Pa], [RS], [S], and [SR].

Of particular interest for us is the work of Padron [P], who considered the equation

$$v_t = \varphi(v)_{xx} + \tau v_{xxt}, \tag{1.10}$$

arising in population dynamics. The point is that φ behaves as in our case, and thus (1.10) is the nondegenerate version of our problem if we replace v by u_x. He proved, with techniques different from ours, an existence theorem and a result which suggests the stability of delta-function-type solutions (which correspond to our piecewise constant solutions).

DiBenedetto and Pierre [dBP] have studied a special class of degenerate pseudoparabolic equations which contains the equation

$$(v - \tau\Delta(\varphi(v)))_t = \Delta(\varphi(v)) \tag{1.11}$$

if φ is nondecreasing. Equations of this form arise in the mathematical models of capillary imbibition and were considered by Barenblatt and Gilman [BG]. Similar equations arise in the theory of fluid flows in fissured or fractured porous media [BZK], [BER], [BS], [Pa], [RS]. We observe that (1.11) can also be viewed as the Yoshida regularization of the (possibly degenerate) equation $v_t = \Delta(\varphi(v))$.

Several papers in the literature are devoted to the maximum principle for pseudoparabolic equations [dBP], [S], [SR], [T]. In this context we observe that many of our proofs are based on maximum principle techniques.

Novick-Cohen and Pego [NP] (see also [N]) studied (1.10) (in arbitrary space dimension) as a model for phase-separation in viscous binary mixtures, but in this case the function φ has a cubic type of nonlinearity. The same sort of nonlinearity arises in the context of viscoelasticity, but in this case the derivative with respect to time is of the second order: (1.7), with u_t replaced by u_{tt} and ψ a linear function, has been studied in [D], [AB], [A], [Pe], and [BHJPS].

2. A brief discussion of the model. For the sake of simplicity we shall only consider the case of a thermally stratified fluid; the case of salinity (density) stratification is obtained in a completely analogous way.

It is known (for references we refer to [BBdPPU], where a more detailed physical discussion of the model is presented) that the mean potential temperature $u(x,t)$ in a statistically horizontally homogeneous turbulent shear flow satisfies the equation

$$\rho c_p u_t + \Phi_x = 0. \tag{2.1}$$

Here t is the time, x is the vertical coordinate, Φ is the turbulent heat flux (i.e., we only take into account heat diffusion due to turbulence, which is supposed to be the dominant diffusion process), and ρ and c_p are, respectively, the fluid density and its heat capacity at constant pressure: fluid properties which are assumed constant. Defining the effective temperature diffusivity k by the relation

$$k = -\frac{\Phi}{\rho c_p u_x}, \tag{2.2}$$

we can rewrite (2.1) in the form

$$u_t = (k u_x)_x. \tag{2.3}$$

Under equilibrium (steady state) conditions the turbulent temperature diffusivity k is a function of the temperature gradient:

$$k = k_0(u_x). \tag{2.4}$$

In turbulent shear flows in stably thermally stratified fluids the temperature gradient u_x is positive, and the main characteristic of such flows is that positive temperature gradients suppress turbulence (this effect is caused by the action of buoyancy forces, due to the presence of a strong gravity field). Therefore (see Fig. 3)

$$k_0'(p) < 0 \quad \text{for } p > 0$$

and

$$k_0(p) \to 0 \quad \text{as } p \to +\infty.$$

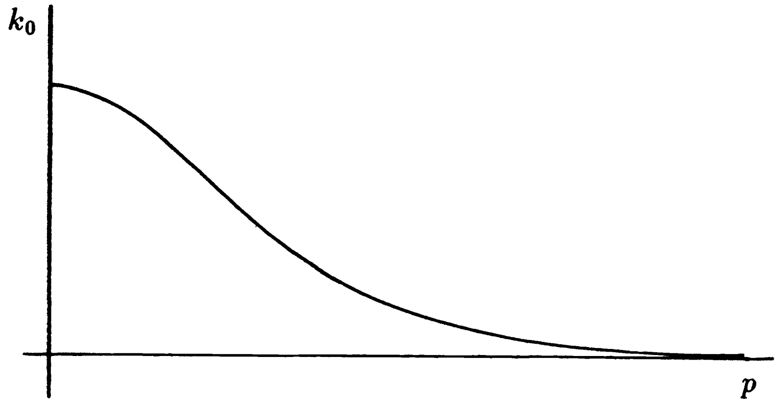


FIG. 3. *The steady heat diffusivity* $k_0(p)$.

Previous investigators have observed that, formally, a sufficiently fast decay of $k_0(p)$ as $p \to +\infty$ may explain the formation of stepwise temperature distributions, a phenomenon which, for example, has been observed in the ocean. Their approach, similar to the traditional derivation of the Fourier heat conduction equation and the Fick diffusion equation, was to use relation (2.4) for closing (2.3), which leads to the following equation for the potential temperature:

$$u_t = \varphi'(u_x)u_{xx}, \tag{2.5}$$

where

$$\varphi(p) = pk_0(p) \quad \text{for } p \geq 0. \tag{2.6}$$

An essential point is that at large values of p the function $k_0(p)$ decreases so rapidly that the function $\varphi(p)$ at a certain value $p = \alpha$ starts to decrease, and, moreover, tends to zero as $p \to +\infty$. Thus, the graph of the function $\varphi(p)$ has the shape which is indicated in Fig. 1 (in the simplest case φ is increasing in $(0, \alpha)$ and decreasing in $(\alpha, +\infty)$), and, for sufficiently large temperature gradients $u_x > \alpha$ the coefficient in (2.5) becomes negative and (2.5) becomes a nonlinear backward heat conduction equation.

Intuitively the negativity of φ' for large values of the temperature gradient might explain the formation of stepwise temperature distributions, and in this direction some numerical experiments were performed [Po], [Dj]. However, as we mentioned in the Introduction, it is plausible that the nonlinear backward-forward heat equation (2.5) leads to ill-posed problems, and therefore the mathematical model should be modified.

To do so we have to take into account that (2.5) is only one of the equations of a system of simultaneous equations governing the turbulence, velocity, and temperature fields. At this moment a generally accepted system of equations to describe these fields is not available. However, it is clear that if in such system the value of the temperature gradient is prescribed at a certain moment, the time τ of adjustment (relaxation) of the turbulence field to this value of the temperature gradient is perhaps small in comparison with the characteristic time of the whole field, but strictly positive. In the present model we average the relaxation time τ over the whole field and consider it as a constant, and we assume that moment t turbulence properties, including the temperature diffusivity k, can be taken as equilibrium values related to the temperature gradient at the moment $t-\tau$.

So, our hypothesis is that the current temperature diffusivity corresponds to the equilibrium temperature gradient at the moment $t-\tau$:

$$k(x,t) = k_0(u_x(x,t-\tau)). \tag{2.7}$$

Bearing in mind that the relaxation time τ is small in comparison with the characteristic time scale of the temperature field we obtain, developing (2.7) in a linear expansion with respect to τ:

$$u_x(x,t-\tau) \approx u_x(x,t) - \tau u_{xt}(x,t)$$

and

$$k \approx k_0(u_x - \tau u_{xt}) \approx k_0(u_x) - \tau k_0'(u_x)u_{xt}. \tag{2.8}$$

Combining (2.5) and (2.8), we obtain the following equation for the temperature:

$$u_t = \varphi(u_x)_x + \tau\psi(u_x)_{xt}, \tag{2.9}$$

where the function $\psi : [0,+\infty) \to [0+\infty)$ is defined by

$$\psi(p) = -\int_0^p s k_0'(s)ds = -\varphi(p) + \int_0^p k_0(s)ds \quad \text{for } p \geq 0. \tag{2.10}$$

Since k_0 is strictly decreasing, ψ is strictly increasing in $\mathbf{R}^+$.

In addition we shall assume that

$$\int_0^{+\infty} k_0(s)ds < +\infty,$$

i.e., that k_0 tends to zero sufficiently fast as $p \to +\infty$, which implies that

$$\psi(+\infty) = \gamma < +\infty.$$

A typical graph of ψ is shown in Fig. 2.

We conclude this section with some comments on the hypotheses which we shall use in the mathematical analysis of the model. It follows from (2.10) that $\psi'(0)=0$ if $pk_0'(p) \to 0$ as $p \to 0$, which makes (2.9) degenerate pseudoparabolic at points at which $u_x=0$. In the rest of this paper we shall ignore this degeneracy at zero. We expect that this only leads to a mathematical complication, and we have preferred to focus our attention to what happens for large gradients, for which φ' becomes negative. In addition, for small values of p the precise behavior of $k_0(p)$ is impossible to determine from experimental data, because the measured quantity is the flux and to obtain the diffusivity we have to divide the flux by the gradient. We observe that, once we assume that $\psi'(p) > 0$ for all p and that $\varphi'(p) < 0$ for large values of p, condition (1.9) implies that $|\varphi'(p)| \le k_1\psi'(p)$ for all p and for some constant k_1. Finally we mention that we shall use an additional condition on φ and ψ (see hypothesis (H2a) in §3) to prove the uniqueness of solutions. In the physical model this condition is satisfied for large gradients if, for example, k_0 is given by (1.3).

3. Basic mathematical hypotheses and main results. We use standard notation for the following function spaces (Ω denotes a connected set in $\mathbf{R}^n$): $C(\Omega)$, $C^k(\Omega)$ and $C^\infty(\Omega)$ denote the sets of, respectively, continuous real functions on Ω, k times differentiable real functions on Ω such that $f^{(k)}$ is continuous in Ω, and real functions on Ω which have derivatives of arbitrary order; $L^p(\Omega)$ $(1 \le p < \infty)$ is the set of measurable real functions on Ω such that $|f|^p$ is Lebesgue integrable in Ω; $L^\infty(\Omega)$ is the set of real measurable functions on Ω which are (essentially) bounded in Ω; $H_0^1(\Omega)$ indicates the Sobolev space which, roughly speaking, contains the functions $u \in L^2(\Omega)$ such that the (generalized) first-order derivatives belong to $L^2(\Omega)$ and u vanishes at the boundary of Ω (in the sense of traces); $BV(\Omega)$ is the set of real functions on Ω which have bounded total variation; $L^p(0,T;V)$ (with V a Banach space with norm $\|\cdot\|$) denotes the set of functions $f : [0,T] \to V$ such that $\|f(t)\|$ belongs to $L^p(0,T)$. Supplied with their natural norms, these sets are Banach spaces ($C(\Omega)$, $C^k(\Omega)$ and $C^\infty(\Omega)$ are Banach spaces if Ω is closed and bounded). The subscript *loc* stands for locally; for example, $L^p_{\text{loc}}(\Omega)$ contains the functions u belonging to $L^p(K)$ for any compact subset K of Ω, and $f_k \to f$ in $L^p_{\text{loc}}(\Omega)$ as $k \to \infty$ indicates that $f_k \to f$ in $L^p(K)$ as $k \to \infty$ for every compact subset K of Ω.

First we list the several hypotheses on the data of problem (I), which we shall use in this paper. From a mathematical point of view it turns out to be convenient to define the functions $\varphi(p)$ and $\psi(p)$ also for negative values of p.

(H1) $\psi \in C^3(\mathbf{R})$, $\psi' > 0$ in $\mathbf{R}$, ψ is odd, and $\psi(+\infty) = \gamma$, where $0<\gamma<+\infty$;

(H2) $\varphi \in C^3(\mathbf{R})$, $\varphi(0) = 0$, and there exists a constant $k_1 > 0$ such that

$$|\varphi'| \le k_1\psi' \quad \text{in } \mathbf{R}; \tag{3.1}$$

(H2a) there exists a constant k_2 such that

$$\left| \left(\frac{\varphi'}{\psi'}\right)' \frac{1}{\psi'} \right| \le k_2 \quad \text{in } \mathbf{R}; \tag{3.2}$$

(H3) $u_0 \in BV((0,1))$, and there exists a function $w_0 \in H_0^1((0,1))$ such that

$$\begin{aligned} w_0(x) &= \lim_{h\to 0} \psi\left(\frac{u_0(x+h)-u_0(x^+)}{h}\right) \\ &= \lim_{h\to 0} \psi\left(\frac{u_0(x+h)-u_0(x^-)}{h}\right) \quad \text{for } 0<x<1, \end{aligned} \tag{3.3}$$

where $u_0(x^{\pm})$ denote the one-sided limits of u_0 at the point x;

(A2) $\varphi(+\infty)=0$, $0<\varphi(p)\le\varphi(\alpha)$ for $0<p<+\infty$ for some $\alpha>0$, and $\varphi<0$ in $\mathbf{R}^-$;

(A3) u_0 is nondecreasing in $(0,1)$.

The hypotheses (H1), (H2), and (H3) will ensure the existence of a solution. The more restrictive condition (H2a) on the behavior of $\psi(p)$ as $p\to\infty$ will be used to prove uniqueness; we observe that it is satisfied in the interval $[1,+\infty)$ in the model case in which φ and ψ are determined by (1.3), (1.4), and (2.10). The assumptions (A2) and (A3) are natural in the context of the physical model, and will be used to study the qualitative behavior of the solutions.

Hypotheses (H1) and (H2) imply that $\varphi(+\infty)$ and $\varphi(-\infty)$ exist and are finite. Defining the function $h:[-\gamma,\gamma]\to\mathbf{R}$ by

$$h(p):=\begin{cases}\varphi\circ\psi^{-1}(p) & \text{if } |p|<\gamma,\\ \varphi(\pm\infty) & \text{if } p=\pm\gamma,\end{cases} \tag{3.4}$$

where ψ^{-1} denotes the inverse of ψ and $\varphi\circ\psi^{-1}$ is the composed function $p\mapsto\varphi(\psi^{-1}(p))$, it follows from (H2) and (H2a) that $h\in C^2((-\gamma,\gamma))\cap C^1([-\gamma,\gamma])$ and

$$|h'|\le k_1 \quad\text{and}\quad |h''|\le k_2 \quad\text{in } (-\gamma,\gamma), \tag{3.5}$$

where we have used the formulas

$$h'(p)=\frac{\varphi'}{\psi'}\circ\psi^{-1}(p) \quad\text{and}\quad h''(p)=\left(\frac{\varphi'}{\psi'}\right)'\frac{1}{\psi'}\circ\psi^{-1}(p) \quad\text{for } -\gamma<p<\gamma.$$

We observe that the continuous function $w_0(x)$ in hypothesis (H3) is nothing else than $\psi(u_0'(x))$ at the points where $|u_0'|<\infty$. In a similar way we replace the function $\psi(u_x(x,t))$, which is only defined at points where $u(x,t)$ is smooth, by a continuous function $w(x,t)$ in the definition of a solution. Throughout this paper we shall use the notation

$$Q=(0,1)\times(0,+\infty) \quad\text{and}\quad Q_T=(0,1)\times(0,T],$$

where $T>0$.

DEFINITION. A function $u:\overline{Q}\mapsto\mathbf{R}$ is a solution of problem (I) if for any $T>0$:

(i) $u\in L^\infty(Q_T)\cap L^\infty(0,T;BV((0,1)))$ and $u_t\in L^2(Q_T)$;

(ii) there exists a function $w\in C(\overline{Q})$ such that $|w|\le\gamma$ in Q,

$$\begin{aligned} w(x,t) &= \lim_{h\to0}\psi\left(\frac{u(x+h,t)-u(x^+,t)}{h}\right)\\ &= \lim_{h\to0}\psi\left(\frac{u(x+h,t)-u(x^-,t)}{h}\right) \quad\text{for } 0<x<1 \quad\text{and}\quad t>0,\end{aligned} \tag{3.6}$$

$w\in L^\infty(0,T;H_0^1((0,1)))$ and $w_t\in L^2(0,T;H^1((0,1)))$;

(iii) u and w satisfy the equation

$$u_t=h(w)_x+\tau w_{xt} \quad\text{in } L^2(Q_T), \tag{3.7}$$

where $h:[-\gamma,\gamma]\to\mathbf{R}$ is defined by (3.4), and

$$u(x,0)+\tau w_x(x,0)=u_0(x)+\tau w_0'(x) \quad\text{for almost every } 0<x<1, \tag{3.8}$$

where u_0 and w_0 are determined by (H3).

In (3.8), the words "almost every" indicate that (3.8) holds in the interval $(0,1)$ with the possible exception of a subset of measure zero.

Our first main result concerns the well-posedness of problem (I).

THEOREM 3.1 (existence and uniqueness). *Let hypotheses* (H1), (H2), *and* (H3) *be satisfied. Then problem* (I) *possesses a solution u. If in addition hypothesis* (H2a) *is satisfied, then the solution is unique.*

In Theorem 3.1 we have not required the positivity of u_0'. On the other hand the positivity of u_x was a crucial physical hypothesis in the derivation of the mathematical model. The first part of the following result shows the consistency of this hypothesis.

THEOREM 3.2 (gradient estimates). *Let hypotheses* (H1)–(H3) *and* A2 *be satisfied, and let $u(x,t)$ be the solution of problem* (I).

(i) *If u_0 is nondecreasing in $(0,1)$, then we have $u(x,t)$ nondecreasing with respect to x.*

(ii) *If $u_x(x_0,t_0) \leq \alpha$ for some $x_0 \in (0,1)$ and $t_0 \geq 0$, then $u_x(x_0,t) \leq \alpha$ for all $t > t_0$.*

(iii) *If $u_x(x_0,t_0) \geq \alpha$ for some $x_0 \in (0,1)$ and $t_0 > 0$, then $u_x(x_0,t) \geq \alpha$ for all $0 \leq t < t_0$.*

The interpretation of part (ii) is the following: $\varphi(p)$ has its maximum at $p=\alpha$, and, in the model equation, φ is increasing in $(0,\alpha)$ and decreasing in $(\alpha,+\infty)$, i.e., the diffusion is positive at points where $0 \leq u_x < \alpha$ and negative where $u_x > \alpha$. As we shall see below, the negative diffusion may lead to unboundedness of the spatial gradient, but Theorem 3.2(ii) means that once $u_x \leq \alpha$ at some point x_0, then it remains bounded by α at x_0 for all later times.

Hypothesis (H3) allows discontinuous initial functions. It is a consequence of the following result that the discontinuities only can increase as time evolves.

THEOREM 3.3 (persistence of discontinuities). *Let hypotheses* (H1)–(H3) *and* (A2)–(A3) *be satisfied. Let u be the solution of problem* (I), *and let w be defined by* (3.6). *If for some $0 < x_0 < 1$ and $t_0 \geq 0$,*

$$w(x_0,t_0) = \gamma,$$

then

$$w(x_0,t) = \gamma \quad \textit{for } t > t_0,$$

and

(3.9) $u(x_0^+,t)$ *is nondecreasing in* $(t_0,+\infty), u(x_0^-,t)$ *is nonincreasing in* $(t_0,+\infty)$.

It turns out that also smooth initial functions may lead to discontinuous solutions.

THEOREM 3.4 (formation of discontinuities). *Let hypotheses* (H1)–(H2a) *and* A2 *be satisfied. Then there exist initial functions $u_0 \in C^1([0,1])$ which satisfy hypotheses* (H3) *and* (A3), *such that the corresponding solutions of problem* (I) *are not continuous in Q.*

Finally we consider the large-time behavior of solutions.

THEOREM 3.5 (convergence to stepwise solutions). *Let hypotheses* (H1)–(H3) *and* (A2)–(A3) *be satisfied and let u be the solution of problem* (I). *Then there exists a nondecreasing function $q \in BV((0,1))$ which satisfies*

$$q' = 0 \quad \textit{almost everywhere in } (0,1)$$

such that

$$u(x,t) \to q(x) \quad \textit{as } t \to \infty \textit{ for almost every } 0 < x < 1. \tag{3.10}$$

If $u \notin C(Q)$, then q is nonconstant in $(0,1)$.

4. Uniqueness. In this section we prove the second part of Theorem 3.1, the uniqueness of the solution of problem (I).

Suppose that $\underline{u}$ and $\overline{u}$ are two solutions with corresponding functions $\underline{w}$ and $\overline{w}$ defined by (3.6). It is sufficient to prove local uniqueness, i.e., $\underline{u} = \overline{u}$ almost everywhere in Q_T for some $T > 0$.

The idea of the proof is to subtract the equations for $(\underline{u}, \underline{w})$ and $(\overline{u}, \overline{w})$ and to multiply the resulting equation by a suitable test function in order to obtain an appropriate integral estimate for the difference of the two solutions.

Let $T > 0$ be determined below. We define $\zeta \in L^2(Q_T)$ by

$$\zeta(x,t) = \int_t^T (\underline{w} - \overline{w})_x(x,s)ds,$$

where $(\underline{w} - \overline{w})_x(x,s)$ stands for $\underline{w}_x(x,s) - \overline{w}_x(x,s)$. Subtracting the equations (3.7) for, respectively, $\underline{u}$ and $\overline{u}$, multiplying by ζ, integrating by parts over Q_T, and using (3.8) and the fact that $\zeta(x,T) = 0$ for almost every $0<x<1$, we obtain that

$$\begin{aligned}\iint_{Q_T} (\underline{u} - \overline{u})(\underline{w} - \overline{w})_x dxdt =& \tau \iint_{Q_T} (\underline{w} - \overline{w})_x^2 dxdt \\ &+ \iint_{Q_T} (h(\underline{w}) - h(\overline{w}))_x \left(\int_t^T (\underline{w} - \overline{w})_x(s,t)ds \right) dxdt.\end{aligned}$$

Splitting up the measure $\underline{u}_x$ into its regular part $\underline{u}_x^R \in L^1(Q_T)$ and its singular part $\underline{u}_x^S$ (see [R]), it follows from the nonnegativity of $\iint_{Q_T} (\underline{u}_x^S - \overline{u}_x^S)(\underline{w} - \overline{w})$ that

$$\begin{aligned}&\tau \iint_{Q_T} (\underline{w} - \overline{w})_x^2 dxdt + \iint_{Q_T} (\underline{u}_x^R - \overline{u}_x^R)(\underline{w} - \overline{w}) dxdt \\ &\le - \iint_{Q_T} (h(\underline{w}) - h(\overline{w}))_x \left(\int_t^T (\underline{w} - \overline{w})_x(s,t)ds \right) dxdt \\ &\le \frac{\epsilon}{2} \iint_{Q_T} (h(\underline{w}) - h(\overline{w}))_x^2 \, dxdt + \frac{1}{2\epsilon} \iint_{Q_T} \left(\int_t^T (\underline{w} - \overline{w})_x(s,t)ds \right)^2 dxdt \\ &\equiv I_1 + I_2,\end{aligned} \tag{4.1}$$

where $\epsilon > 0$ is a constant to be chosen below. First we estimate I_1:

$$\begin{aligned}I_1 \le& \epsilon \iint_{Q_T} \left((h'(\underline{w})(\underline{w} - \overline{w})_x)^2 + (h'(\underline{w}) - h'(\overline{w}))^2 \overline{w}_x^2 \right) dxdt \\ \le& \epsilon k_1^2 \iint_{Q_T} (\underline{w} - \overline{w})_x^2 dxdt + \epsilon k_2^2 \sup_{[0,T]} \left(\int_0^1 \overline{w}_x^2(x,t)dx \right) \\ &\cdot \int_0^T \sup_{(0,1)} (\underline{w} - \overline{w})^2 dt,\end{aligned} \tag{4.2}$$

where we have used (3.5). Observe that, for any $t \in [0,T]$,

$$\begin{aligned}\sup_{0<x<1} (\underline{w} - \overline{w})^2(x,t) &= 2 \sup_{0<x<1} \int_0^x (\underline{w} - \overline{w})(\underline{w} - \overline{w})_x(s,t)ds \\ &\le \int_0^1 (\underline{w} - \overline{w})^2(x,t)dx + \int_0^1 (\underline{w} - \overline{w})_x^2(x,t)dx.\end{aligned}$$

We use the second term on the left-hand side of (4.1) to estimate $\underline{w}-\overline{w}$ in $L^2(Q_T)$: since $\underline{w}=\psi(\underline{u}_x^R)$ almost everywhere in Q_T, we have

$$\iint_{Q_T}(\underline{u}_x^R-\overline{u}_x^R)(\underline{w}-\overline{w})dxdt \ge \frac{1}{\sup_{\mathbf{R}}\psi'}\iint_{Q_T}(\underline{w}-\overline{w})^2dxdt,$$

and hence it follows from (4.1) and (4.2) that if we choose $\epsilon>0$ such that

$$\epsilon k_1^2 \le \frac{\tau}{4} \quad\text{and}\quad \epsilon k_2^2\left(\sup_{[0,1]}\int_0^1 \overline{w}_x^2(x,t)dx\right) \le \max\left\{\frac{\tau}{4},\frac{1}{2\sup_{\mathbf{R}}\psi'}\right\},$$

then, for any $T\in(0,1]$,

$$\begin{aligned}
&\frac{1}{2}\tau\iint_{Q_T}(\underline{w}-\overline{w})_x^2dxdt+\frac{1}{2\sup_{\mathbf{R}}\psi'}\iint_{Q_T}(\underline{w}-\overline{w})^2dxdt\\
&\le\frac{1}{2\epsilon}\iint_{Q_T}\left(\int_t^T(\underline{w}-\overline{w})_x(x,s)ds\right)^2dxdt\\
&\le\frac{1}{2\epsilon}\iint_{Q_T}(T-t)\left(\int_t^T(\underline{w}-\overline{w})_x^2ds\right)dxdt\le\frac{T^2}{2\epsilon}\iint_{Q_T}(\underline{w}-\overline{w})_x^2dxdt.
\end{aligned}$$

Choosing $T\in(0,1]$ such that $T^2\le\frac{1}{2}\epsilon\tau$, we find that

$$\frac{1}{4}\tau\iint_{Q_T}(\underline{w}-\overline{w})_x^2dxdt+\frac{1}{2\sup_{\mathbf{R}}\psi'}\iint_{Q_T}(\underline{w}-\overline{w})^2dxdt\le 0,$$

and the proof is complete.

5. Existence. To prove the existence of a solution of problem (I), we proceed in two steps. The first one consists of approximating problem (I) by a nondegenerate problem.

For any $0<\epsilon\le1$ there exist functions φ_ϵ, $\psi_\epsilon\in C^\infty(\mathbf{R})$, and $u_{0\epsilon}\in C^\infty([0,1])$ such that

$$\begin{gathered}
\psi'+\epsilon\le\psi_\epsilon'\le\psi'+2\epsilon \quad\text{in }\mathbf{R},\quad \psi_\epsilon''\in L^\infty(\mathbf{R}),\\
\psi_\epsilon\text{ is odd},\quad \psi_\epsilon\to\psi\quad\text{in }C^3_{\text{loc}}(\mathbf{R})\quad\text{as }\epsilon\to0,
\end{gathered}$$

$$\varphi_\epsilon(0)=0,\quad |\varphi_\epsilon'|\le k_1\psi_\epsilon'\quad\text{in }\mathbf{R},\quad \varphi_\epsilon\to\varphi\quad\text{in }C^3_{\text{loc}}(\mathbf{R})\quad\text{as }\epsilon\to0,$$

where k_1 is defined by (3.1), and

$$\begin{gathered}
\int_0^1\Psi_\epsilon(u_{0\epsilon}'(x))dx\le C,\qquad \int_0^1\left(\psi_\epsilon(u_{0\epsilon}'(x))'\right)^2dx\le C,\\
u_{0\epsilon}'(0)=u_{0\epsilon}'(1)=0,\quad u_{0\epsilon}\to u_0\quad\text{in }L^1(0,1)\quad\text{as }\epsilon\to0
\end{gathered}\tag{5.1}$$

for some $C>0$, which does not depend on ϵ, where we have set

$$\Psi_\epsilon(p)=\int_0^p\psi_\epsilon(s)ds\quad\text{for }p\in\mathbf{R}.\tag{5.2}$$

If condition (A2) is satisfied, we may suppose that

$$\varphi_\epsilon < 0 \quad \text{in } \mathbf{R}^-, \quad 0 \leq \varphi_\epsilon(p) \leq \varphi_\epsilon(\alpha) \quad \text{in } \mathbf{R}^+,$$

and

$$\varphi_\epsilon = 0 \quad \text{in } (\psi_\epsilon^{-1}(\gamma), +\infty). \tag{5.3}$$

We consider the approximate problem

$$(\mathrm{I}_\epsilon)\begin{cases} u_t = \varphi_\epsilon(u_x)_x + \tau\psi_\epsilon(u_x)_{xt} & \text{in } Q, \\ u(x,0) = u_{0\epsilon}(x) & \text{for } 0 < x < 1, \\ u_x(0,t) = u_x(1,t) = 0 & \text{for } t > 0. \end{cases}$$

LEMMA 5.1. *Let $\epsilon > 0$ and $k = 0, 1, 2, \ldots$. Then problem (I_ϵ) has a unique solution*

$$u_\epsilon \in W^{1,\infty}_{\text{loc}}([0,+\infty); C^{2+k}([0,1])).$$

(The notation $u_\epsilon \in W^{1,\infty}_{\text{loc}}$ indicates that, if we set $Y = C^{2+k}([0,1])$ with norm $\|\cdot\|$, the functions $u_\epsilon(t)$ and $u_{\epsilon t}(t)$ belong to Y for all $t \geq 0$ and the scalar functions $\|u_\epsilon(t)\|$ and $\|u_{\epsilon t}(t)\|$ belong to $L^\infty(0,T)$ for any $T > 0$.)

Proof. We introduce the spaces

$$X = \{u \in C^{2+k}([0,1]) : u'(0) = u'(1) = 0\},$$
$$Y = C^k([0,1]),$$

and we define the functionals $G : X \to Y$ and $H : X \to Y$ by

$$G(u) = u - \tau\psi_\epsilon(u')' \quad \text{and} \quad H(u) = \varphi_\epsilon(u')'. \tag{5.4}$$

Clearly these functionals are uniformly Lipschitz continuous. In addition we claim that G is invertible and that its inverse

$$G^{-1} : Y \to X \quad \text{is locally Lipschitz continuous.} \tag{5.5}$$

First, we prove (5.5) for $k=0$.

Let $v_1, v_2 \in Y$ and consider the problem

$$(\mathrm{II}_{\epsilon,i})\begin{cases} u_i - \tau\psi_\epsilon(u_i')' = v_i & \text{in } (0,1), \\ u_i'(0) = u_i'(1) = 0, \end{cases}$$

where $i = 1, 2$. By standard results on quasilinear elliptic equations and by the maximum principle, there exists a unique solution $u_i \in X$ of problem $(\mathrm{II}_{\epsilon,i})$, which satisfies

$$\max_{[0,1]} |u_i| \leq \max_{[0,1]} |v_i| \quad \text{and} \quad \|u_i\|_X \leq C\|v_i\|_Y \tag{5.6}$$

for some constant C that does not depend on v_i.

To prove (5.5), we consider the equation for $z := u_1 - u_2$:

$$z - \tau\psi'_\epsilon(u'_1)z'' - \tau\left(\psi'_\epsilon(u'_1) - \psi'_\epsilon(u'_2)\right)u''_2 = v_1 - v_2,$$

i.e.,

$$z - \tau\psi'_\epsilon(u'_1)z'' - \tau\psi''_\epsilon(\xi)u''_2 z' = v_1 - v_2,$$

where $\xi(x)$ is a number between $u'_1(x)$ and $u'_2(x)$. By the maximum principle

$$\|z\|_Y \leq \|v_1 - v_2\|_Y,$$

and thus

$$|\psi'_\epsilon(u'_1)z'' + \psi''_\epsilon(\xi)u''_2 z'| \leq \frac{2}{\tau}\|v_1 - v_2\|_Y,$$

from which (5.5) follows at once for $k=0$.

If $k \geq 1$, (5.6) remains valid. The remaining part of the proof of (5.5) is based on a standard iteration procedure applied to the equation for the kth-order derivative of $u_1 - u_2$.

We rewrite problem (I_ϵ) as an ordinary differential equation in Y:

$$\begin{cases} \dfrac{d}{dt}w = (H \circ G^{-1})(w) & \text{in } [0, +\infty), \\ w(0) = G(u_{0\epsilon}). \end{cases}$$

The local existence of a unique solution $w_\epsilon \in C^1([0,T];Y)$ (for some $T > 0$) follows from (5.5), and by (5.6) the solution can be extended globally: $w_\epsilon \in C^1([0,+\infty);Y)$. Applying (5.5) a second time, we find that

$$u_\epsilon := G^{-1}(w_\epsilon) \in W^{1,\infty}_{\mathrm{loc}}([0,\infty);X),$$

and u_ϵ is the unique solution of problem (I_ϵ).

The second step in the existence proof is to show several integral estimates for u_ϵ which will enable us to let $\epsilon \to 0$.

LEMMA 5.2. *Let u_ϵ be the solution of problem* (I_ϵ), *and let Ψ_ϵ be defined by* (5.2). *For any $T > 0$ there exists a constant C which does not depend on ϵ such that*

$$\sup_{t\in[0,T]} \int_0^1 \Psi_\epsilon(u_{\epsilon x})dx \leq C, \tag{5.7}$$

$$\sup_{t\in[0,T]} \int_0^1 \psi_\epsilon(u_{\epsilon x})^2_x dx \leq C, \tag{5.8}$$

$$\iint_{Q_T} u^2_{\epsilon t} dxdt \leq C, \tag{5.9}$$

$$\iint_{Q_T} \psi'_\epsilon(u_{\epsilon x})u^2_{\epsilon xt} dxdt \leq C, \tag{5.10}$$

$$\iint_{Q_T} \psi_\epsilon(u_{\epsilon x})^2_{xt} dxdt \leq C, \tag{5.11}$$

$$\sup_{t\in[0,T]} \int_0^1 \psi_\epsilon(u_{\epsilon x})^2 dx \leq C, \tag{5.12}$$

$$\int_0^1 u_\epsilon(x,t)dx = \int_0^1 u_{0\epsilon}(x)dx. \tag{5.13}$$

Proof. Multiplying the equation for u_ϵ by $\psi_\epsilon(u_{\epsilon x})_x$ and integrating by parts, we obtain that for $t \in (0,T]$

$$\int_0^1 \left(\Psi_\epsilon(u_{\epsilon x}) + \frac{1}{2}\tau\psi_\epsilon(u_{\epsilon x})_x^2\right)(x,t)dx - \int_0^1 \left(\Psi_\epsilon(u'_{0\epsilon}) + \frac{1}{2}\tau(\psi_\epsilon(u'_{0\epsilon})')^2\right)(x)dx$$
$$= -\iint_{Q_t} \psi_\epsilon(u_{\epsilon x})_x\varphi_\epsilon(u_{\epsilon x})_x dxdt \le k_1 \iint_{Q_t} \psi_\epsilon(u_{\epsilon x})_x^2 dxdt,$$

where k_1 is defined by (3.1). It follows easily from the conditions on ψ_ϵ and $u_{0\epsilon}$ that the second integral on the left-hand side is uniformly bounded with respect to ϵ, and (5.7) and (5.8) follow at once.

We multiply the equation for u_ϵ by $u_{\epsilon t}$ and integrate by parts:

$$\iint_{Q_T} u_{\epsilon t}^2 dxdt + \tau \iint_{Q_T} \psi'_\epsilon(u_{\epsilon x})u_{\epsilon xt}^2 dxdt = -\int_0^1 \Phi_\epsilon(u_{\epsilon x})(x,T)dx + \int_0^1 \Phi_\epsilon(u'_{0\epsilon}(x))dx,$$

where we have set

$$\Phi_\epsilon(p) = \int_0^p \varphi(s)ds \quad \text{for } p \in \mathbf{R}.$$

Since $|\Phi_\epsilon| \le C\Psi_\epsilon$ in $\mathbf{R}$ for some constant C independent of ϵ, the right-hand side is uniformly bounded, and we obtain (5.9) and (5.10).

Finally (5.11) follows from (5.8), (5.9), and the equation for u_ϵ, (5.12) follows from (5.7) and the inequality $\psi_\epsilon^2 \le C\Psi_\epsilon$ for some C independent of ϵ, and integration of the equation for u_ϵ implies (5.13).

Remark. Except for the estimate (5.11), all estimates of Lemma 5.2 can be obtained if we use, instead of (3.1), the weaker inequality $\varphi' \ge -k_1\psi'$ (cf. (1.9)). In particular it can be shown that this condition is sufficient to prove existence of a solution in the sense of distributions.

Now we are ready to construct a solution u of problem (I). It follows from (5.7) and (5.9) that the set $\{u_\epsilon\}_{0<\epsilon\le 1}$ is bounded in $BV_{\text{loc}}(\overline{Q})$, and it follows from a straightforward diagonal procedure that there exist $u \in L^1_{\text{loc}}(\overline{Q})$ and a sequence $\{\epsilon_n\}$ converging to 0 such that

$$u_{\epsilon_n} \to u \quad \text{in } L^1_{\text{loc}}(\overline{Q}) \quad \text{as } n \to \infty.$$

We observe that (5.13) implies that

$$\int_0^1 u(x,t)dx = \int_0^1 u_0(x)dx \quad \text{for } t \ge 0.$$

Since $u \in L^\infty(0,T; BV((0,1)))$, this implies that $u \in L^\infty(Q_T)$.

In view of (5.8), (5.10)–(5.12), we may assume that there exists $w \in L^\infty_{\text{loc}}((0,+\infty); H^1_0(0,1))$ such that $w_t \in L^2_{\text{loc}}((0,+\infty); H^1(0,1))$ and

$$\psi_{\epsilon_n}(u_{\epsilon_n x}) \to w \quad \text{weakly in } H^1_{\text{loc}}(\overline{Q}) \quad \text{as } n \to \infty.$$

It follows easily from the integral estimates (5.8), (5.10), and (5.11) for $\psi_\epsilon(u_{\epsilon x})$ that these functions are uniformly Hölder continuous with exponent $\frac{1}{2}$ in Q_T; thus $w \in C^{1/2}(Q_T)$ and we may assume that

$$\psi_{\epsilon_n}(u_{\epsilon_n x}) \to w \quad \text{uniformly in } Q_T \quad \text{as } n \to \infty.$$

In addition $|w| \leq \gamma$ in Q_T. Indeed, if $|w| > \gamma + 2\eta$ at some point $(x_0, t_0) \in Q_T$ for some $\eta > 0$, then there exists a neighborhood $U \subset Q_T$ of (x_0, t_0) which does not depend on n such that for all sufficiently large n,

$$|\psi_{\epsilon_n}(u_{\epsilon_n})| > \gamma + \eta \quad \text{in } U;$$

this implies that

$$|u_{\epsilon_n}| > \psi_{\epsilon_n}^{-1}(\gamma + \eta) \quad \text{in } U,$$

and since the term at the right-hand side tends to infinity as $n \to \infty$ we obtain a contradiction with the boundedness of u_{ϵ_n} in $BV_{\mathrm{loc}}(\overline{Q})$.

Property (3.6) follows immediately from the construction and continuity of w.

Using the equation for u_ϵ we obtain for any $\zeta \in C^2(\overline{Q}_T)$ the integral identity

$$\int_0^1 (u\zeta + \tau w\zeta_x)(x, T)dx - \int_0^1 (u_0(x)\zeta(x, 0) + \tau w_0(x)\zeta_x(x, 0))dx$$
$$= \iint_{Q_T} (u\zeta_t + w\zeta_{xt} - h(w)\zeta_x)dxdt,$$

i.e., u and w satisfy (3.7) in the sense of distributions. In view of the regularity of u and w, (3.7) is satisfied in $L^2(Q_T)$.

Next we prove a local regularity result for the limiting function w which we shall use in §8.

THEOREM 5.3. *Let hypotheses* (H1), (H2), *and* (H3) *be satisfied, let* w_0 *be defined as in* (H3), *and let* u *be the solution of problem* (I) *constructed in this section, with the corresponding function* w *defined by* (3.6). *Let* K *be a closed interval contained in* $\{x \in [0,1] : |w_0(x)| < \gamma\}$. *If* $w_0 \in H^2(K)$, *then for every* $\bar{t} > 0$ *such that* $K \times [0, \bar{t}] \subset \{(x,t) : |w(x,t)| < \gamma\}$,

$$w \in L^\infty((0,\bar{t}); H^2(K)) \cap C^{1+\frac{1}{2}}(K \times [0,\bar{t}]).$$

(Here $H^2(K)$ is the Sobolev space containing the functions $u \in L^2(K)$ such that the generalized derivatives of u of first- and second-order belong to $L^2(K)$.)

Proof. Let $\{u_{0\epsilon}\}$ be a sequence of $C^\infty([0,1])$-functions, equibounded in $H^2(K)$, such that $u_{0\epsilon} \to u_0$ in $L^1([0,1])$ as $\epsilon \to 0$. Let u_ϵ be the corresponding solution of problem (I_ϵ), and $w_\epsilon = \psi_\epsilon(u_{\epsilon x}) \to w$ in $C(\overline{Q_T})$ as $\epsilon \to 0$. Thus there exist $\epsilon' > 0$ and $\gamma' \in (0, \gamma)$ such that $|w_\epsilon| \leq \gamma' < \gamma$ in $K \times [0, \bar{t}]$ for every $0 < \epsilon < \epsilon'$. It follows easily from (5.9) and (5.10) that

$$\int_0^{\bar{t}} u_{\epsilon t}^2(x,t)dt \leq \tilde{C}(\gamma') \quad \text{for all } x \in K, \tag{5.14}$$

where $\tilde{C}(\gamma')$ does not depend on $\epsilon \in (0, \epsilon')$.

Multiplying the equation for u_ϵ by $\psi_\epsilon(u_{\epsilon x})_x$ and integrating over $[0, t^*]$, where $t^* \leq \bar{t}$, we obtain for every $x \in K$,

$$\frac{1}{2}\tau\psi_\epsilon(u_{\epsilon x})_x^2(x, t^*) \leq \frac{1}{2}\tau\psi_\epsilon(u_{\epsilon x})_x^2(x, 0) + \left(\frac{1}{2} + k_1\right)\int_0^{t^*} \psi_\epsilon(u_{\epsilon x})_x^2(x,t)dt$$
$$+ \frac{1}{2}\int_0^{\bar{t}} u_{\epsilon t}^2(x,t)dt,$$

where k_1 is defined by (3.1). Combining this integral inequality with (5.14), we find that the functions

$$\psi_\epsilon(u_{\epsilon x})_x \ (0<\epsilon<\epsilon') \quad \text{are uniformly bounded in } K \times [0,\bar{t}]. \tag{5.15}$$

From (5.14) and (5.15) it follows that $\psi_\epsilon(u_{\epsilon x})_x$ are Hölder-continuous with respect to t with exponent $\frac{1}{2}$ uniformly with respect to $x \in K$; indeed, it turns out that

$$\begin{aligned}
&\tau \int_0^{t^*} \psi_\epsilon(u_{\epsilon x})^2_{xt}(x,t)dt \\
&\quad \le \int_0^{t^*} |\varphi_\epsilon(u_{\epsilon x})_x \psi_\epsilon(u_{\epsilon x})_{xt}|(x,t)dt + \int_0^{t^*} |u_{\epsilon t}\psi_\epsilon(u_{\epsilon x})_{xt}|(x,t)dt \\
&\quad \le \frac{1}{2}\tau \int_0^{t^*} \psi_\epsilon(u_{\epsilon x}(x,t))^2_{xt}dt + \frac{1}{\tau}\left(\int_0^{t^*} (\varphi_\epsilon(u_{\epsilon x}(x,t))_x)^2 dt + \int_0^{t^*} (u_{\epsilon t}(x,t))^2 dt\right).
\end{aligned}$$

To complete the proof we derive an estimate for $\psi_\epsilon(u_{\epsilon x})_{xx}$ in $L^2(K)$. For this purpose we consider the equation for $v_\epsilon = u_{\epsilon x}$:

$$v_{\epsilon t} = \varphi_\epsilon(v_\epsilon)_{xx} + \tau\psi_\epsilon(v_\epsilon)_{xxt}. \tag{5.16}$$

Multiplying (5.16) by $\psi_\epsilon(v_\epsilon)_{xx}$ and integrating by parts over $K \times [0,t^*]$, where $0<t^*<\bar{t}$, we have

$$\begin{aligned}
&\tau \int_0^{t^*} \frac{d}{dt}\left(\int_K \tfrac{1}{2}\psi_\epsilon(v_\epsilon)^2_{xx}dx\right) dt \\
&\quad = -\iint_{K\times[0,t^*]} \varphi_\epsilon(v_\epsilon)_{xx}\psi_\epsilon(v_\epsilon)_{xx}dxdt + \iint_{K\times[0,t^*]} v_{\epsilon t}\psi_\epsilon(v_\epsilon)_{xx}dxdt \\
&\quad \le c_1 \iint_{K\times[0,t^*]} \psi_\epsilon(v_\epsilon)^2_{xx}dxdt + c_2 \iint_{K\times[0,\bar{t}]} \psi_\epsilon(v_\epsilon)^4_x dxdt \\
&\qquad + c_3 \iint_{K\times[0,\bar{t}]} \psi'_\epsilon(v_\epsilon)v^2_{\epsilon t}dxdt \\
&\quad \le c_1 \iint_{K\times[0,t^*]} \psi_\epsilon(v_\epsilon)^2_{xx}dxdt + c_4,
\end{aligned}$$

where we have used (5.15), (5.10), and the fact that $\psi'_\epsilon(v_\epsilon)$ is uniformly bounded away from zero in K. It follows at once from this inequality that

$$\psi_\epsilon(v_\epsilon)_{xx} \in L^\infty([0,\bar{t}], L^2(K)),$$

and the proof is complete.

6. Gradient extimates. In this section we shall prove the estimates for u_x given in Theorem 3.2.

Proof of Theorem 3.2(i). It is sufficient to prove that if $u'_{0\epsilon} \ge 0$ in $(0,1)$, the solution u_ϵ of problem (I_ϵ) satisfies the inequality

$$u_{\epsilon x} \ge 0 \quad \text{in } Q_T.$$

The proof is inspired by a method used by Novick-Cohen and Pego [NP]. Let $g : \mathbf{R} \to \mathbf{R}$ be a smooth nondecreasing function such that

$$g(s)\begin{cases} = 0 & \text{if } s \geq 0, \\ < 0 & \text{if } s < 0. \end{cases}$$

Then

$$K_\epsilon(p) := \int_0^p g(\varphi_\epsilon(s))ds\begin{cases} = 0 & \text{if } p \geq 0, \\ > 0 & \text{if } p < 0, \end{cases}$$

and in particular,

$$\int_0^1 K_\epsilon(u'_{0\epsilon})dx = 0.$$

Hence for all $0 < t \leq T$,

$$\begin{aligned}\int_0^1 K_\epsilon(u_{\epsilon x})(x,t)dx &= \iint_{Q_t} g(\varphi_\epsilon(u_{\epsilon x}))u_{\epsilon xt}dxdt \\ &= -\iint_{Q_t} \left(g(\varphi_\epsilon(u_{\epsilon x}) + \tau\psi_\epsilon(u_{\epsilon x})_t) - g(\varphi_\epsilon(u_{\epsilon x}))\right) u_{\epsilon xt}dxdt \\ &\quad + \iint_{Q_t} g(\varphi_\epsilon(u_{\epsilon x}) + \tau\psi_\epsilon(u_{\epsilon x})_t)u_{\epsilon xt}dxdt.\end{aligned}$$

By the mean value theorem, the first term at the right-hand side is nonpositive. Integrating the second integral by parts and using the equation for u_ϵ, we obtain

$$\int_0^1 K_\epsilon(u_{\epsilon x})(x,t)dx \leq -\iint_{Q_t} g'(\varphi_\epsilon(u_{\epsilon x}) + \tau\psi_\epsilon(u_{\epsilon x})_t)u_{\epsilon t}^2 dxdt \leq 0.$$

This implies that $K_\epsilon(u_{\epsilon x}) = 0$ in Q_T and thus $u_{\epsilon x} \geq 0$ in Q_T.

Proof of Theorem 3.2(ii)–(iii). We set

$$z_\epsilon(x,t) = \int_0^x u_\epsilon(y,t)dy, \qquad 0 \leq x \leq 1, t \geq 0, \tag{6.1}$$

where u_ϵ is the solution of problem (I_ϵ). Then

$$z_{\epsilon x} = u_\epsilon \quad \text{in } Q \tag{6.2}$$

and

$$z_{\epsilon t} = \varphi_\epsilon(u_{\epsilon x}) + \tau\psi_\epsilon(u_{\epsilon x})_t \quad \text{in } Q. \tag{6.3}$$

Substituting (6.2) into the latter term of (6.3), we have

$$z_{\epsilon t} - \tau\psi'_\epsilon(u_{\epsilon x})z_{\epsilon txx} = \varphi_\epsilon(u_{\epsilon x}) \quad \text{in } Q. \tag{6.4}$$

By (6.3),

$$z_{\epsilon t}(0,t) = z_{\epsilon t}(1,t) = 0 \quad \text{for } t \geq 0; \tag{6.5}$$

hence, applying for fixed t the maximum principle to the ordinary differential equation (6.4) in x for $z_{\epsilon t}$, we find that

$$z_{\epsilon t} \leq \max_{x\in[0,1]} \varphi_\epsilon(u_{\epsilon x})(x,t) \leq \varphi_\epsilon(\alpha) \quad \text{in } Q.$$

In view of (6.3) this means that

$$\tau\psi_\epsilon(u_{\epsilon x})_t \leq \varphi_\epsilon(\alpha) - \varphi_\epsilon(u_{\epsilon x}) \quad \text{in } Q. \tag{6.6}$$

Fixing $x = x_0$ in (6.6) we obtain an ordinary differential inequality for $u_{\epsilon x}(x_0, t)$, which is uniform with respect to ϵ near the value $u_{\epsilon x} = \alpha$, and Theorem 3.2(ii)–(iii) follows at once.

7. The persistence of discontinuities. In this section we prove that if $u_x(x_0, t_0) = +\infty$ for some (x_0, t_0), then $u_x(x_0, t)$ remains infinite for all $t > 0$. The main idea of the proof is to show that we can reconstruct the solution for $t > t_0$ by solving the problem independently in the sets $(0, x_0) \times (t_0, +\infty)$ and $(x_0, 1) \times (t_0, +\infty)$ if we impose the boundary condition $u_x(x_0, t) = +\infty$ for $t > 0$. This construction reflects the lack of interaction between these two sets across the point $x = x_0$, caused by the strong degeneracy of the equation for infinite gradients; in other words, the vertical line $x = x_0$ is a characteristic of the equation.

Proof of Theorem 3.3. Without loss of generality we may assume that $t_0 = 0$. We consider the problem

$$(\mathrm{III}_\epsilon)\begin{cases} u_t = \varphi_\epsilon(u_x)_x + \tau\psi_\epsilon(u_x)_{xt} & \text{in } (x_0, 1) \times \mathbf{R}^+, \\ u(x, 0) = u_{0\epsilon}(x) & \text{for } x \in (x_0, 1), \\ \psi_\epsilon(u_x(x_0, t)) = \gamma & \text{for } t > 0, \\ \psi_\epsilon(u_x(1, t)) = 0 & \text{for } t > 0. \end{cases}$$

Arguing as in the proof of Theorem 3.1, it follows that problem (III_ϵ) has a unique solution u_ϵ, and that there exists a sequence $\epsilon_n \to 0$ as $n \to \infty$ such that u_{ϵ_n} converges to a solution u of the problem

$$(\mathrm{III})\begin{cases} u_t = \varphi(u_x)_x + \tau\psi(u_x)_{xt} & \text{in } (x_0, 1) \times \mathbf{R}^+, \\ u(x, 0) = u_0(x) & \text{for } x_0 < x < 1, \\ \psi(u_x(x_0, t)) = \gamma & \text{for } t > 0, \\ \psi(u_x(1, t)) = 0 & \text{for } t > 0. \end{cases}$$

(The definition of a solution of problem (III) is quite similar to the one of problem (I).)

We claim that the solution of problem (III) has the following property:

$$(7.1) \qquad u(x_0^+, t) \quad \text{is nondecreasing for } t > 0.$$

Before proving this we complete the proof of Theorem 3.3.

In a similar way we solve a corresponding problem for $0 < x < x_0$:

$$(\mathrm{III}')\begin{cases} u_t = \varphi(u_x)_x + \tau\psi(u_x)_{xt} & \text{in } (0, x_0) \times \mathbf{R}^+, \\ u(x, 0) = u_0(x) & \text{for } 0 < x < x_0, \\ \psi(u_x(x_0, t)) = \gamma & \text{for } t > 0, \\ \psi(u_x(0, t)) = 0 & \text{for } t > 0, \end{cases}$$

and its solution $\overline{u}(x, t)$ satisfies the property that

$$(7.2) \qquad \overline{u}(x_0^-, t) \quad \text{is nonincreasing in } (0, T].$$

Finally, defining

$$\hat{u}(x, t) = \begin{cases} u(x, t) & \text{if } x_0 < x \le 1, t \ge 0, \\ \overline{u}(x, t) & \text{if } 0 \le x < x_0, t \ge 0, \end{cases}$$

it is easy to show that $\hat{u}$ is the unique solution of problem (I), and the desired result follows from (7.1) and (7.2).

So it remains to prove (7.1).

Let

$$v_\epsilon(x,t) = \int_{x_0}^{x} u_\epsilon(y,t)dy \quad \text{if } x_0 \le x \le 1, t \ge 0 \tag{7.3}$$

and

$$v(x,t) = \int_{x_0}^{x} u(y,t)dy \quad \text{if } x_0 \le x \le 1, t \ge 0. \tag{7.4}$$

Arguing as in the proof of Theorem 3.2(ii)–(iii) and using the boundary conditions at $x=x_0$ and $x=1$, we obtain that $v_{\epsilon t}(\cdot,t)$ satisfies for any $t>0$,

$$\begin{cases} v_{\epsilon t} - \tau\psi_\epsilon'(u_{\epsilon x})v_{\epsilon txx} = \varphi_\epsilon(u_{\epsilon x}) \ge 0 & \text{for } x_0 < x < 1, \\ v_{\epsilon t}(x_0,t) = v_{\epsilon t}(1,t) = 0. \end{cases}$$

Hence $v_\epsilon(x,t)$ and $v(x,t)$ are nondecreasing with respect to t. Since $v(x_0,t)=0$ for $t>0$, this implies that

$$v_x(x_0,t) = u(x_0+,t) \quad \text{is nondecreasing for } t>0,$$

and we have found (7.1).

8. The formation of discontinuities. In this section we shall prove that smooth solutions may develop discontinuities within finite time. The idea of the proof is to start off with a suitable initial function which is discontinuous at some point x_0, to solve the problem backwards in time, and to show that the solution becomes smooth after a finite (negative) time.

Proof of Theorem 3.4. Let $x_0 \in (0, \frac{1}{2})$, and let a, c and δ be small positive parameters such that $c > a\delta$. Then there exists a function u_0 (see Fig. 4) satisfying hypotheses (H3) and (A3) such that

$$\begin{aligned} &u_0 \in C^3([0,1] \setminus \{x_0, 1-x_0\}), \\ &u_0(x_0+) = a, \quad u_0(x_0-) = 0, \quad u(x_0+\delta) = a+c, \\ &u_0(1-x) - u_0(\tfrac{1}{2}) = u_0(\tfrac{1}{2}) - u_0(x) \quad \text{for almost every } 0 < x < 1, \end{aligned}$$

$$u'(x)\begin{cases} > \alpha & \text{if } x_0 < x < x_0+\delta, \\ = \alpha & \text{if } x_0+\delta \le x \le \frac{1}{2}, \end{cases}$$

and

$$u_0''(x) \ge \beta \quad \text{for } 0 \le x < x_0$$

for some $\beta > 0$.

Using the transformation $t \to -t$ and $\varphi(s) \to -\varphi(s)$, we may apply Theorem 3.1 to solve problem (I) backwards in time, i.e., problem (I) has a unique solution $u(x,t)$ for $t<0$. By uniqueness, u has the symmetry of its initial function:

$$u(1-x,t) - u_0(\tfrac{1}{2}) = u_0(\tfrac{1}{2}) - u(x,t) \quad \text{for almost every } 0 < x < 1, t < 0. \tag{8.1}$$

In addition, it easily follows from the continuity of the function w defined by (3.6) and from Theorem 5.3 that there exists a $\underline{t}_0 < 0$ such that $u(x,t)$ is strictly increasing with respect to x in $(0,1) \times [\underline{t}_0, 0)$.

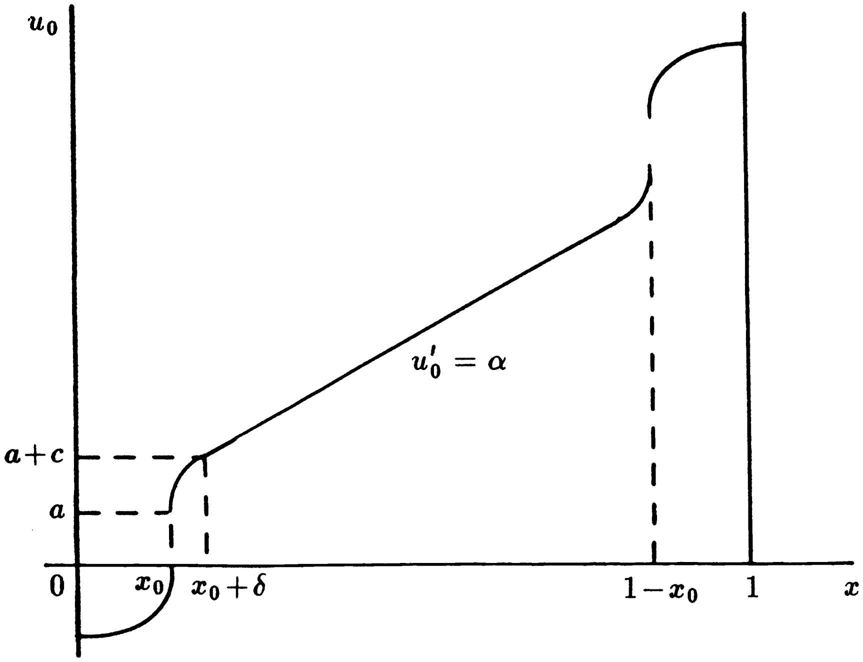


FIG. 4. *The function* u_0 *in the proof of Theorem* 3.4.

We claim that for an appropriate choice of the parameters a, c, and δ, there exists $t_0 \in (\underline{t}_0, 0)$ such that u is smooth at (x_0, t_0). By symmetry, u is also smooth at $(1-x_0, t_0)$; on the other hand it follows from Theorem 3.3 and the smoothness of u_0 that $|u_x(x,t_0)| < +\infty$ in the remaining points (x, t_0). Hence the continuous function $\psi(u_x)$ satisfies $|\psi(u_x)(x,t_0)| < \gamma$ for all $0 \le x \le 1$, and $u(\cdot, t_0) \in C^1([0,1])$. Using $u(\cdot, t_0)$ as the initial function of problem (I) forwards in time, we obtain Theorem 3.4.

It remains to prove our claim. Arguing by contradiction, we may suppose that the function $w(x,t)$, defined in(3.6), satisfies

$$w(x_0,t) = w(1-x_0,t) = \gamma \quad \text{for all } t < 0.$$

Hence the function $u(x,t)$, restricted to the set $[x_0, 1-x_0] \times (-\infty, 0]$, is a solution of the problem

$$\text{(IV)}\begin{cases} u_t = \varphi(u_x)_x + \tau\psi(u_x)_{xt} & \text{for } x_0 < x < 1-x_0, t < 0, \\ \psi(u_x(x_0,t)) = \psi(u_x(1-x_0,t)) = \gamma & \text{for } t < 0, \\ u(x,0) = u_0(x) & \text{for } x_0 < x < 1-x_0. \end{cases}$$

It is not difficult to adapt the proof in §5 to show the uniqueness of the solution of problem (IV).

It follows from Theorem 3.2(iii) that

$$\text{(8.2)} \qquad u_x(x,t) \ge \alpha \quad \text{for almost every } x_0 < x < 1-x_0, t < 0,$$

and Theorem 3.3 implies that

$$\text{(8.3)} \qquad u(x_0+,t) \ge u(x_0-,t) \ge u_0(x_0-) = 0$$

and

$$u((1-x_0)-,t) \le u((1-x_0)+,t) \le u_0((1-x_0)+) \le 2(a+c)+\alpha(1-2x_0). \tag{8.4}$$

Combining (8.2), (8.3), and (8.4) we have that

$$\alpha(x-x_0) \le u(x,t) \le 2(a+c)+\alpha(x-x_0) \quad \text{for almost every } x_0 < x < 1-x_0, t<0. \tag{8.5}$$

The solution of problem (IV) may be approximated by the solution u_ϵ of the problem

$$(\mathrm{IV}_\epsilon)\begin{cases} u_t = \varphi_\epsilon(u_x)_x + \tau\psi_\epsilon(u_x)_{xt} & \text{for } x_0 < x < 1-x_0, t<0, \\ \psi_\epsilon(u_x(x_0,t)) = \psi_\epsilon(u_x(1-x_0,t)) = \gamma & \text{for } t<0, \\ u(x,0) = u_{0\epsilon}(x) & \text{for } x_0 < x < 1-x_0, \end{cases}$$

where $u_{0\epsilon}$ is a smooth function which converges uniformly to u_0 in $(x_0, 1-x_0)$ such that $u'_{0\epsilon} \ge \alpha$ in $(x_0, 1-x_0)$.

Let $v_\epsilon(x,t)$ and $v(x,t)$ be defined by (7.3) and (7.4) for $x_0 \le x \le 1-x_0$ and $t \le 0$. By (8.5) we have that for all $x_0 \le x \le 1-x_0$ and $t \le 0$

$$v(x,t)-v(x,0) = \int_{x_0}^{x} (u(s,t)-u_0(s))ds \ge -2(a+c).$$

Since $v_\epsilon \to v$ uniformly on bounded sets as $\epsilon \to 0$, there exists $\epsilon_0 > 0$ such that for all $0<\epsilon<\epsilon_0$,

$$v_\epsilon(x,t)-v_\epsilon(x,0) \ge -3(a+c) \quad \text{for } x_0 \le x \le 1-x_0, -1 \le t \le 0. \tag{8.6}$$

Arguing as in §7, we obtain that

$$v_{\epsilon t} = \tau\psi_\epsilon(u_{\epsilon x})_t + \varphi_\epsilon(u_{\epsilon x}) \quad \text{in } (x_0, 1-x_0) \times (-\infty, 0).$$

Hence

$$v_\epsilon(x,t)-v_\epsilon(x,0) = \tau\psi_\epsilon(u_{\epsilon x}(x,t)) - \tau\psi_\epsilon(u'_{0\epsilon}(x)) - \int_t^0 \varphi_\epsilon(u_{\epsilon x}(x,s))ds,$$

and, by (8.6), for $0<\epsilon<\epsilon_0$

$$\begin{aligned} \tau\psi_\epsilon(u_{\epsilon x}(x,t)) \ge & \tau\psi_\epsilon(\alpha) + \int_t^0 \varphi_\epsilon(u_{\epsilon x}(x,s))ds - 3(a+c) \\ & \text{for } x_0 \le x \le 1-x_0, -1 \le t \le 0. \end{aligned} \tag{8.7}$$

Since $\varphi_\epsilon(s)$ is uniformly bounded away from zero in a neighborhood of $s=\alpha$ for ϵ small enough, it follows from (8.7) that there exist constants $0<\epsilon_1 \le \epsilon_0$, $\underline{t}_0 \le t_0 < 0$ and $C>0$ which do not depend on a, c, and δ such that

$$u_{\epsilon x}(x,t_0) \ge C+\alpha-3(a+c) \quad \text{for } x_0 < x < 1-x_0.$$

Letting $\epsilon \to 0$, we obtain that

$$u_x(x,t_0) \ge C+\alpha-3(a+c) \quad \text{for almost every } x_0 < x < 1-x_0,$$

and, choosing $a+c$ sufficiently small, we obtain a contradiction with (8.5).

9. The convergence to stepwise solutions. In this section we prove the convergence of the solution to a stepwise steady-state as $t \to +\infty$. The first step is rather elementary: integrating the equation with respect to x, we shall use the maximum principle to prove the existence of an asymptotic profile.

More precisely, let z_ϵ be defined by (6.1), and let

$$z(x,t) = \int_0^x u(y,t)dy \quad \text{for } 0 \le x \le 1, t \ge 0.$$

Since $\varphi_\epsilon(u_{\epsilon x}) \ge 0$ in Q, it follows from (6.4) and (6.5) that $z_{\epsilon t} \ge 0$ in Q, and hence

$$z(x,t) \quad \text{is nondecreasing with respect to } t. \tag{9.1}$$

Since u is nondecreasing with respect to x,

$$z(x,t) \quad \text{is convex with respect to } x, \tag{9.2}$$

and since $z(0,t)=0$ and $z(1,t)=\int_0^1 u_0(x)dx$ for all $t\ge 0$, it follows from (9.1) and (9.2) that the pointwise limit

$$\overline{z}(x) = \lim_{t\to\infty} z(x,t), \qquad 0 \le x \le 1$$

exists, and

$$\overline{z}(x) \quad \text{is convex in } [0,1].$$

Since $u = z_x$ is uniformly bounded in $[0,1]$,

$$\overline{z} \in W^{1,\infty}((0,1)),$$

and it follows from (9.2) that

$$u(x,t) \to q(x) \equiv \overline{z}'(x) \quad \text{as } t \to \infty \quad \text{for almost every } 0 < x < 1. \tag{9.3}$$

To complete the proof of Theorem 3.5 we have to prove that

$$q'(x) = 0 \quad \text{for almost every } 0 < x < 1, \tag{9.4}$$

which is the second step of the proof. The fact that $q(x)$ is nonconstant if u has a discontinuity at some point (x_0, t_0) follows at once from (3.9).

We observe that, formally, (9.4) follows if we prove that $\varphi(q(x))=0$ for all x, i.e., that $q(x)$ is a solution of the steady-state problem

$$\begin{cases} \varphi(q')' = 0 & \text{in } (0,1), \\ q'(0) = q'(1) = 0. \end{cases}$$

Below we shall make this argument precise.

Integrating (6.3) with respect to t, we obtain that for any $t_1 > t_0 \ge 0$ and $0 < x < 1$,

$$z_\epsilon(x,t_1) - \tau\psi_\epsilon(u_{\epsilon x}(x,t_1)) = z_\epsilon(x,t_0) - \tau\psi_\epsilon(u_{\epsilon x}(x,t_0)) + \int_{t_0}^{t_1} \varphi_\epsilon(u_{\epsilon x}(x,t))dt.$$

Since $z_\epsilon \to z$ and $\psi_\epsilon(u_{\epsilon x}) \to w$ uniformly on bounded sets as $\epsilon \to 0$, we have that for any $t_1 > t_0 \geq 0$ and $0 < x < 1$,

$$z(x,t_1) - \tau w(x,t_1) = z(x,t_0) - \tau w(x,t_0) + \int_{t_0}^{t_1} h(w(x,t))dt, \tag{9.5}$$

where h is defined by (3.4).

Since $h(w) \geq 0$, $0 \leq w \leq \gamma$, and $z \to \overline{z}$ as $t \to \infty$, we may define

$$\overline{w}(x) = \lim_{t\to\infty} w(x,t) \quad \text{for } 0 \leq x \leq 1.$$

We observe that this implies that

$$w(\cdot,t) \to \overline{w} \quad \text{in } L^1((0,1)) \quad \text{as } t \to \infty, \tag{9.6}$$

and

$$\int_0^\infty h(w(x,s))ds \quad \text{is uniformly bounded in } (0,1).$$

In particular, since for all $0<x<1$ and $t>0$,

$$\int_t^{t+1} h(w(x,s))ds = z(x,t+1) - z(x,t) - \tau w(x,t+1) + \tau w(x,t),$$

we find in the limit $t \to \infty$ that

$$h(\overline{w}(x)) = 0 \quad \text{for all } 0 < x < 1,$$

which implies that the function $\overline{w}$ assumes only the values 0 or γ.

Let $\epsilon > 0$. Because of (9.6), and since $z(\cdot,t) \to \overline{z}$ uniformly in $(0,1)$, there exists $t_\epsilon > 0$ such that $t_\epsilon \to \infty$ as $\epsilon \to 0$,

$$0 \leq \overline{z} - z(\cdot,t) < \epsilon\tau \quad \text{in } (0,1) \quad \text{for } t \geq t_\epsilon, \tag{9.7}$$

and

$$\text{measure } \{x \in (0,1) : |w(x,t) - \overline{w}(x)| \geq \epsilon\} < \epsilon \quad \text{for } t \geq t_\epsilon. \tag{9.8}$$

We assume for the moment that

$$\overline{w} = 0 \quad \text{almost everywhere in } (0,1). \tag{9.9}$$

Defining the open set

$$U_\epsilon = \{x \in (0,1) : w(x,t_\epsilon) < \epsilon\},$$

it follows from (9.8) and (9.9) that

$$\text{measure } U_\epsilon > 1 - \epsilon.$$

We claim that

$$q' \leq \psi^{-1}(2\epsilon) \quad \text{in } U_\epsilon. \tag{9.10}$$

In view of the arbitrariness of ϵ, this completes the proof of (9.4).

It follows from (9.5) that for all $t > t_\epsilon$ and $0 \le x \le 1$,

$$\tau w(x,t) = \tau w(x,t_\epsilon) + (z(x,t) - \overline{z}(x)) - (z(x,t_\epsilon) - \overline{z}) - \int_{t_\epsilon}^{t} h(w(x,s))ds,$$

and thus, by (9.7), the positivity of $h(w)$ and the definition of U_ϵ,

$$w(x,t) \le 2\epsilon \quad \text{in } U_\epsilon \quad \text{for } t > t_\epsilon.$$

It follows from the relation $w = \psi(u_x)$ in the set where $w < \gamma$ that

$$u_x(x,t) < \psi^{-1}(2\epsilon) \quad \text{for } x \in U_\epsilon \quad \text{and} \quad t > t_\epsilon.$$

Since U_ϵ is an open subset of $(0,1)$, this implies (9.10).

It remains to prove (9.9). We argue by contradiction, and suppose that the set

$$S = \{x \in (0,1) : \overline{w}(x) = \gamma\}$$

has positive measure. Since

$$\tau w(x,t) = \tau \overline{w}(x) + z(x,t) - \overline{z}(x) + \int_{t}^{\infty} h(w(x,s))ds$$

and $z(\cdot,t) \to \overline{z}$ uniformly in $(0,1)$ as $t \to \infty$, for any $\delta > 0$ there exists a $t_\delta > 0$ such that

$$w(x,t) \ge \gamma - \delta \quad \text{for } x \in S \quad \text{and} \quad t > t_\delta.$$

Since

$$w(x,t) = \lim_{\epsilon \to 0} \psi_\epsilon(u_{\epsilon x}) \quad \text{uniformly on bounded sets,}$$

a straightforward integration of $u_{\epsilon x}$ yields that

$$u(1,t) - u(0,t) \ge \psi^{-1}(\gamma - \delta)|S|.$$

Since the choice of $\delta > 0$ is arbitrary and $\psi^{-1}(\gamma - \delta) \to \infty$ as $\delta \to 0$, we obtain a contradiction with the boundedness of u.

To conclude the paper, we observe that it follows easily from Theorem 3.3 about the persistence of discontinuities, that the (infinitely many) piecewise constant steady-states are all locally stable (for example, in $L^\infty(0,1)$ or in $BV(0,1)$), and therefore the asymptotic profile will strongly depend on the initial function u_0.

We conjecture that the asymptotic profile also depends on the value of the relaxation time τ. More precisely, we guess that, given an initial function which has supercritical gradient in some subinterval, the number of discontinuities of the corresponding asymptotic profile may become arbitrarily large as $\tau \to 0$. Some numerical evidence for this conjecture can be found in [BBdPPU].

Acknowledgments. The authors wish to express their sincere thanks to the Institute of Mathematics and its Applications, University of Minnesota, Minneapolis, where this paper was started. The coordinating role of Professor Shoshana Kamin is warmly appreciated.

REFERENCES

[A] G. ANDREWS, *On the existence of solutions to the equation* $u_{tt} = u_{xxt} + \sigma(u_x)_x$, J. Differential Equations, 35 (1980), pp. 200–231.

[AB] G. ANDREWS AND J. M. BALL, *Asymptotic behaviour and changes of phase in one–dimensional nonlinear viscoelasticity*, J. Differential Equations, 44 (1982), pp. 306–341.

[BHJPS] J. M. BALL, P. J. HOLMES, R. D. JAMES, R. L. PEGO AND P. J. SWART, *On the dynamics of fine structure*, J. Nonlinear Science, 1 (1991), pp. 17–70.

[BBdPPU] G. I. BARENBLATT, M. BERTSCH, R. DAL PASSO, V. M. PROSTOKISHIN, AND M. UGHI, *A mathematical model of turbulent heat and mass transfer in stably stratified shear flow*, J. Fluid Mech., 253 (1993), pp. 341–358.

[BER] G. I. BARENBLATT, V. M. ENTOV, AND V. M. RYZHIK, *Theory of Fluid Flows in Natural Rocks*, Kluwer Academic Publishers, Dordrecht, Boston, London, 1990.

[BG] G. I. BARENBLATT AND A. A. GILMAN, *Mathematical model of non-equilibrium counter current capillary imbibition*, J. Engrg. Phys. (INZH–Phys.), 3 (1987), pp. 456–461.

[BZK] G. I. BARENBLATT, I. ZHELTOV, AND I. KOCHINA, *Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks*, J. Appl. Math. Mech., 24 (1960), pp. 1286–1303.

[BS] M. BÖHM AND R. E. SHOWALTER, *Diffusion in fissured media*, SIAM J. Math. Anal., 16 (1985), pp. 500–509.

[CS] R. W. CARROLL AND R. E. SHOWALTER, *Singular and Degenerate Cauchy Problems*, Academic Press, New York, 1976.

[D] C. M. DAFERMOS, *The mixed initial–boundary value problem for the equations of nonlinear viscoelasticity*, J. Differential Equations, 6 (1969), pp. 71–86.

[dBP] E. DIBENEDETTO AND M. PIERRE, *On the maximum principle for pseudo-parabolic equations*, Indiana Univ. Math. J., 30 (1981), pp. 821–854.

[Dj] S.KH. DJUMAGAZIEVA, *Numerical integration of a certain partial differential equation*, J. Num. Math. Phys., 23 (1983), pp. 839–847.

[H] K. HÖLLIG, *Existence of infinitely many solutions for a forward backward heat equation*, Trans. Amer. Math. Soc., 278 (1983), pp. 299–316.

[N] A. NOVICK–COHEN, *On the viscous Cahn–Hilliard equations*, in Material Instabilities in Continuum Mechanics and Related Mathematical Problems, J.M. Ball, ed., Oxford, 1988.

[NP] A. NOVICK-COHEN AND R.L. PEGO, *Stable patterns in a viscous diffusion equation*, Trans. Amer. Math. Soc., 324 (1991), pp. 331–351.

[P] V. PADRON, *Sobolev Regularization of Some Nonlinear Ill-Posed Problems*, Ph.D. thesis, University of Minnesota, Minneapolis, 1990.

[Pa] H. PASCAL, *On nonlinear effects in unsteady flows of non–newtonian fluids through fractured porous media*, Internat J. Nonlinear Mechanics, 26 (1991), pp. 487–499.

[Pe] R. L. PEGO, *Phase transitions in one–dimensional nonlinear visco–elasticity: admissibility and stability*, Arch. Rational Mech. Anal., 97 (1987), pp. 353–394.

[Po] E. S. POSMENTIER, *The generation of salinity finestructures by vertical diffusion*, J. Phys. Oceanogr., 7–3 (1977), pp. 298–300.

[R] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1976.

[RS] J. RULLA AND S. E. SHOWALTER, *Diffusion with prescribed convection in fissured media*, Differential Integral Equations, 1 (1988), pp. 315–325.

[S] R. E. SHOWALTER, *Local regularity, boundary values and maximum principles for pseudoparabolic equations*, Appl. Anal., 16 (1983), pp. 235–241.

[SR] M. STECHER AND W. RUNDELL, *The nonpositivity of solutions to pseudoparabolic equations*, Proc. Amer. Math. Soc., 75 (1979), pp. 251–254.

[T] T.W. TING, *A cooling process according to two-temperature theory of heat conduction*, J. Math. Anal. Appl., 45 (1974), pp. 23–31.

SIAM J. MATH. ANAL.
Vol. 24, No. 6, pp. 1440–1457, November 1993

003

ON A TRANSMISSION BOUNDARY-VALUE PROBLEM FOR THE TIME-HARMONIC MAXWELL EQUATIONS WITHOUT DISPLACEMENT CURRENTS*

MARTIN REISSEL†

Abstract. This paper considers a transmission boundary-value problem for the time-harmonic Maxwell equations neglecting displacement currents. The usual transmission conditions, which require the continuity of the tangential components of the electric and magnetic fields across boundaries, are slightly modified. For this new problem, it is shown that the uniqueness of the solution depends on the topological properties of the domains under consideration. Finally, existence results are obtained by using a boundary integral equation approach.

Key words. time-harmonic Maxwell equations, existence and uniqueness, integral equation methods

AMS subject classifications. 35Q60, 45F99, 78A25

1. Introduction. Many problems in electrical engineering lead to transmission boundary-value problems for the time-harmonic Maxwell equations. A standard problem of this type is shown in Fig. 1. One considers a bounded domain $G^E \subset \mathbb{R}^3$ of conducting material which is surrounded by an isolator (usually air). In the interior of the unbounded domain $G^L := \mathbb{R}^3 \backslash \bar{G}^E$ a time-harmonic current density $\widetilde{J}_e(x,t) = J_e(x)e^{-i\omega t}$ is given. We are now interested in the currents induced in G^E by $\widetilde{J}_e$. This leads to the classical transmission boundary-value problem for the time-harmonic Maxwell equations

$$
\begin{aligned}
&\left.\begin{aligned}\operatorname{curl} H^L &= J_e - i\omega\varepsilon^L E^L\\ \operatorname{curl} E^L &= i\omega\mu^L H^L\end{aligned}\right\} \quad \text{in } G^L,\\
&\left.\begin{aligned}\operatorname{curl} H^E &= (\sigma^E - i\omega\varepsilon^E)E^E\\ \operatorname{curl} E^E &= i\omega\mu^E H^E\end{aligned}\right\} \quad \text{in } G^E,\\
&\left.\begin{aligned}n \wedge H^L &= n \wedge H^E\\ n \wedge E^L &= n \wedge E^E\end{aligned}\right\} \quad \text{on } \partial G^E,
\end{aligned}
\tag{1}
$$

with the Silver–Müller radiation condition

$$
H^L \wedge \frac{x}{|x|} - E^L = o\left(\frac{1}{|x|}\right)
$$

uniformly for $|x| \to \infty$.

* Received by the editors October 14, 1992; accepted for publication (in revised form) March 22, 1993.

† Fachbereich Mathematik, Universitaet Kaiserslautern, D-6750 Kaiserslautern, Germany.

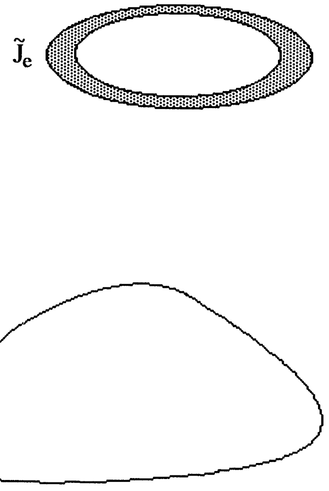


FIG. 1

The different constants have the following meaning:

$\omega \geq 0$	frequency,
$\varepsilon^L, \varepsilon^E > 0$	electric permittivity in G^L, G^E,
$\mu^L, \mu^E > 0$	magnetic permeability in G^L, G^E,
$\sigma^E > 0$	electric conductivity in G^E.

Under certain assumptions on the regularity of J_e and the smoothness of the boundary Γ, which separates the domains G^L and G^E, existence and uniqueness of solutions H^L, E^L, H^E, E^E of (1) can be shown [7], [9].

Dealing with problems in connection with machines working at power frequencies, equations (1) are modified. Since the frequency ω is very small, displacement currents are usually neglected, which means that ε^L and ε^E are set to zero in (1). Moreover, the transmission and radiation conditions are changed. The continuity of the tangential components of the electric field across Γ is substituted by the condition $n \cdot (\mu^L H^L) = n \cdot (\mu^E H^E)$ on Γ, n being the outer normal to G^E. In addition, the Silver–Müller radiation condition is replaced by $H^L(x) = o(1)$, $E^L(x) = o(1)$ uniformly for $|x| \to \infty$. All these modifications together yield our new problem as follows.

$$\left.\begin{aligned} \operatorname{curl} H^L &= J_e \\ \operatorname{curl} E^L &= i\omega\mu^L H^L \end{aligned}\right\} \text{ in } G^L, \qquad \left.\begin{aligned} \operatorname{curl} H^E &= \sigma^E E^E \\ \operatorname{curl} E^E &= i\omega\mu^E H^E \end{aligned}\right\} \text{ in } G^E,$$

$$\text{(2)} \qquad \left.\begin{aligned} n \wedge H^E &= n \wedge H^L \\ n \cdot (\mu^E H^E) &= n \cdot (\mu^L H^L) \end{aligned}\right\} \text{ on } \Gamma,$$

$$H^L(x) = o(1), \qquad E^L(x) = o(1) \quad \text{uniformly for } |x| \to \infty.$$

As we will see, uniqueness results for (2) will strongly depend on the topology of G^L respectively, G^E. In §4 we will show existence for (2), and the set of all possible solutions will be completely characterized.

2. Preliminaries. Before we start with the existence and uniqueness proof, we want to give a detailed description of the problem.

Let $C(G)(C^k(G)$ respectively, $C^{0\alpha}(G)$, $0 < \alpha < 1)$ denote the space of continuous (k times continuously differentiable respectively, Hölder continuous) functions on G.

$G^E \subset \mathbb{R}^3$ is an open, bounded domain with C^2 boundary. The complement $G^L = \mathbb{R}^3 \backslash \bar{G}^E$ should be connected ($\bar{G}^E$ denotes the closure of G^E). G^E is the union of m connected components $G_j^E, j = 1, \dots, m$ having the topological genus p_j. The boundaries $\Gamma_j = \partial G_j^E$ are closed surfaces, which should be disjoint. Setting $\Gamma = \bigcup_{j=1}^m \Gamma_j$ we get $\Gamma = \partial G^E = \partial G^L$.

The topological genus of G^E, respectively G^L, is $p = \sum_{j=1}^m p_j$. There exist p surfaces $\sum_i^E \subset G^E$, respectively, $\sum_i^L \subset G^L$, $i = 1, \dots, p$, such that $G^E \backslash \bigcup_{i=1}^p \sum_i^E$, respectively, $G^L \backslash \bigcup_{i=1}^p \sum_i^L$ are simply connected. The boundary curves $\gamma_i^L = \partial \sum_i^E$ and $\gamma_i^E = \partial \sum_i^L$ lie on Γ.

Example. Let G^E be a torus. In this case we have $m = p = 1$. The surfaces $\sum_1^E, \sum_1^L$ and the curves γ_1^L, γ_1^E are shown in Fig. 2.

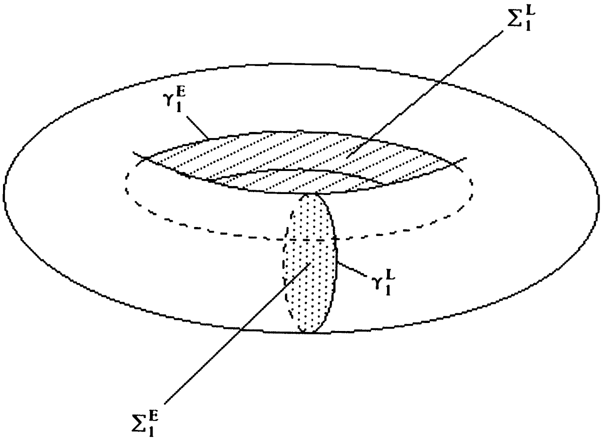


FIG. 2

The problem to be solved is now defined as: for $J_e \in C^1(\mathbb{R}^3)$, div $J_e = 0$, supp$(J_e) \subset G^J$, $\bar{G}^J \subset G^L$ bounded, find $H^L, E^L \in C^1(G^L) \cap C(\bar{G}^L)$, $H^E, E^E \in C^1(G^E) \cap C(\bar{G}^E)$, solving

$$(3) \qquad \begin{array}{c} \operatorname{curl} H^L = J_e \\ \operatorname{curl} E^L = i\omega\mu^L H^L \end{array} \text{ in } G^L, \qquad \begin{array}{c} \operatorname{curl} H^E = \sigma^E E^E \\ \operatorname{curl} E^E = i\omega\mu^E H^E \end{array} \text{ in } G^E,$$

$$\begin{aligned} n \wedge H^E &= n \wedge H^L \\ n \cdot (\mu^E H^E) &= n \cdot (\mu^L H^L) \end{aligned} \qquad \text{on } \Gamma, \tag{4}$$

$$H^L(x) = o(1), \qquad E^L(x) = o(1) \quad \text{uniformly for } |x| \to \infty. \tag{5}$$

3. Uniqueness. Taking a closer look at (3)–(5), it becomes obvious that we cannot expect uniqueness for all four fields because adding the gradient of a suitably chosen function to E^L does not change anything in (3)–(5). Therefore, if we talk about uniqueness in the sequel, we only mean uniqueness of the fields H^L, H^E, and E^E.

THEOREM 1. *For problem* (3)–(5), *together with the additional condition*

$$\int_{\gamma_i^L} \tau \cdot H^L dl = h_i^L, \qquad i = 1, \ldots, p, \tag{6}$$

$h_i^L \in \mathbb{C}$ *given,* τ *being the unit tangent to* γ_i^L, *the fields* H^L, H^E, E^E *are uniquely determined.*

Proof. We consider the homogeneous problem with $J_e = 0$ and $h_i^L = 0, i = 1, \ldots, p$. We show that the fields H^L, H^E, E^E vanish identically.

From the first transmission condition $n \wedge H^E = n \wedge H^L$ on Γ, we get, with the help of the Gaussian theorem,

$$\begin{aligned} \int_\Gamma n \cdot (\bar{H}^L \wedge E^E) ds &= \int_\Gamma n \cdot (\bar{H}^E \wedge E^E) ds \\ &= \int_{G^E} (\sigma^E E^E \cdot \bar{E}^E - i\omega\mu^E H^E \cdot \bar{H}^E) dv, \end{aligned} \tag{7}$$

where $\bar{F}$ denotes the complex conjugate of the field F.

G^E and G^L were defined to have topological genus p. In this case, it is well known [6] that there exist p linear independent Neumann fields Z_i^E respectively, Z_i^L, $i = 1, \ldots, p$, in G^E respectively, G^L, fulfilling

$$\text{curl } Z_i^E = 0, \quad \text{div } Z_i^E = 0 \quad \text{in } G^E, \quad n \cdot Z_i^E = 0 \quad \text{on } \Gamma,$$

$$\text{curl } Z_i^L = 0, \quad \text{div } Z_i^L = 0 \quad \text{in } G^L, \quad n \cdot Z_i^L = 0 \quad \text{on } \Gamma,$$

$$\int_{\gamma_i^L} \tau \cdot Z_j^L dl = \delta_{ij}, \quad \int_{\gamma_i^E} \tau \cdot Z_j^L dl = 0, \quad \int_{\gamma_i^E} \tau \cdot Z_j^E dl = \delta_{ij}, \quad \int_{\gamma_i^L} \tau \cdot Z_j^E dl = 0,$$

and

$$Z_i^L(x) = O\left(\frac{1}{|x|^2}\right),$$

uniformly for $|x| \to \infty$. As a consequence of the regularity assumptions on G^E and G^L we get

$$Z_i^E \in C^\infty(G^E) \cap C^{0\alpha}(\bar{G}^E), \qquad Z_i^L \in C^\infty(G^L) \cap C^{0\alpha}(\bar{G}^L).$$

Using the second transmission condition $n \cdot (\mu^L H^L) = n \cdot (\mu^E H^E)$ on Γ, we conclude that for any surface element $S \subset \Gamma$ we have

$$\int_{\partial S} \tau \cdot (E^E - E^L)dl = \int_S n \cdot \text{curl }(E^E - E^L)ds = i\omega \int_S n \cdot (\mu^E H^E - \mu^L H^L)ds = 0.$$

But this means that the tangential components of $E^E - E^L$ on Γ are of the form

$$(E^E - E^L)\big|_{\text{tan}} = \text{Grad } \varphi + \sum_{j=1}^{p} e_j^L Z_j^L + \sum_{j=1}^{p} e_j^E Z_j^E,$$

where Grad φ denotes the surface gradient of φ on Γ and $e_i^L, e_i^E \in \mathbb{C}, i = 1, \dots, p$, are complex numbers.

In complete analogy, we derive from

$$\text{curl } H^L = 0 \quad \text{in } G^L, \qquad \int_{\gamma_i^L} \tau \cdot H^L dl = 0, \quad i = 1, \dots, p,$$

that we can write the tangential components of H^L as a surface gradient

$$H^L\big|_{\text{tan}} = \text{Grad } \psi.$$

Putting $(E^E - E^L)|_{\text{tan}}$ and $H^L|_{\text{tan}}$ in (7), we arrive at

$$\int_\Gamma n \cdot (\bar{H}^E \wedge E^E)ds = \int_\Gamma n \cdot (\bar{H}^L \wedge E^E)ds$$

$$= \int_\Gamma n \cdot \left(\text{Grad } \bar{\psi} \wedge \left(E^L + \text{Grad } \varphi + \sum_{i=1}^{p} e_i^L Z_i^L + \sum_{i=1}^{p} e_i^E Z_i^E\right)\right) ds.$$

Applying Stokes's theorem to the terms on the right-hand side, we deduce

$$\int_\Gamma n \cdot (\text{Grad } \bar{\psi} \wedge \text{Grad } \varphi)ds = 0,$$

$$\int_\Gamma n \cdot (\text{Grad } \bar{\psi} \wedge Z_i^L)ds = 0,$$

$$\int_\Gamma n \cdot (\text{Grad } \bar{\psi} \wedge Z_i^E)ds = 0,$$

and therefore,

$$\int_\Gamma n \cdot (\bar{H}^E \wedge E^E)ds = \int_\Gamma n \cdot (\bar{H}^L \wedge E^L)ds. \tag{8}$$

Let us now consider $G^R = G^L \cap B^R, B^R := \{x | x \in \mathbb{R}^3, |x| \leq R\}$. For large enough R, we get, by using the Gaussian theorem,

$$\int_{\partial B^R} \frac{x}{|x|} \cdot (\bar{H}^L \wedge E^L)ds - \int_\Gamma n \cdot (\bar{H}^L \wedge E^L)ds = \int_{\partial G^R} n' \cdot (\bar{H}^L \wedge E^L)ds$$

$$= -i\omega\mu^L \int_{G^R} H^L \cdot \bar{H}^L dv,$$

n' being the outer normal to G^R.

Together with (7) and (8) this means

$$\begin{aligned}&-i\omega\mu^L \int_{G^R} H^L \cdot \bar{H}^L dv \\ (9) \qquad &+ \int_{G^E} (\sigma^E E^E \cdot \bar{E}^E - i\omega\mu^E H^E \cdot \bar{H}^E) dv = \int_{\partial B^R} \frac{x}{|x|} \cdot (\bar{H}^L \wedge E^L) ds.\end{aligned}$$

For H^L we have

$$\operatorname{curl} H^L = 0, \quad \operatorname{div} H^L = 0 \quad \text{in } G^L, \quad H^L(x) = o(1) \quad \text{uniformly for } |x| \to \infty.$$

Therefore, H^L is a harmonic vector field in G^L tending to zero for $|x| \to \infty$ and thus [6]

$$H^L(x) = O\left(\frac{1}{|x|^2}\right), \quad \text{uniformly for } |x| \to \infty.$$

From $E^L(x) = o(1)$ uniformly for $|x| \to \infty$ we get

$$\left|\frac{x}{|x|} \cdot (\bar{H}^L \wedge E^L)\right| = o\left(\frac{1}{R^2}\right) \quad \text{on } \partial B^R, \quad R \to \infty$$

and

$$\int_{\partial B^R} \frac{x}{|x|} \cdot (\bar{H}^L \wedge E^L) ds = o(1) \quad \text{for } R \to \infty.$$

Taking the limit $R \to \infty$, (9) yields

$$\sigma^E \int_{G^E} E^E \cdot \bar{E}^E dv - i\omega \left(\mu^E \int_{G^E} H^E \cdot \bar{H}^E dv + \mu^L \int_{G^L} H^L \cdot \bar{H}^L dv\right) = 0.$$

Since $\omega, \sigma^E, \mu^L, \mu^E$ are real and positive constants, we conclude

$$H^L \equiv 0, \quad E^E \equiv 0, \quad H^E \equiv 0. \qquad \square$$

Remark. The free parameters h_i^L in (6), which are the circulations of H^L along the curves γ_i^L, are later on used to characterize the different solutions of (3)–(5).

4. Existence. To establish existence results for (3)–(5) we consider the following auxiliary problem.

Find $H^L \in C^1(G^L) \cap C(\bar{G}^L)$,

$$H^E \in C^2(G^E) \cap C(\bar{G}^E), \quad \operatorname{div} H^E \in C(\bar{G}^E), \quad \operatorname{curl} H^E \in C(\bar{G}^E)$$

solving

$$\left.\begin{aligned}\operatorname{curl} H^L &= 0 \\ \operatorname{div} H^L &= 0\end{aligned}\right\} \text{in } G^L, \qquad \left.\begin{aligned}(\Delta + k^2) H^E &= 0 \\ k^2 &= i\omega\sigma^E\mu^E\end{aligned}\right\} \text{in } G^E,$$

$$(10)\qquad \begin{aligned} n \wedge H^E - n \wedge H^L &= c \\ n \cdot (\mu^E H^E) - n \cdot (\mu^L H^L) &= g \quad \text{on } \Gamma, \\ \operatorname{div} H^E &= d \end{aligned}$$

$$\int_{\gamma_i^L} \tau \cdot H^L dl = 0, \qquad i = 1, \ldots, p,$$

$$H^L(x) = o(1) \quad \text{uniformly for } |x| \to \infty.$$

For k we choose the square root with positive imaginary part.

In this chapter we show that (10) is uniquely solvable for sufficiently smooth data c, g, d. Moreover, we describe how all solutions of (3)–(5) can be constructed by using the solvability of the auxiliary problem (10).

By the following lemma [8], the uniqueness of (10) can be reduced to the uniqueness theorem given in the last chapter.

LEMMA 1. *Let* $H \in C^1(G^L) \cap C(\bar{G}^L)$, $H(x) = O(1/|x|^{1+\beta})$, $|x| \to \infty$, $0 < \beta < 2$, *satisfying* div $H = 0$ *in* G^L. *If we have*

$$\int_{\Gamma_j} n \cdot H \, ds = 0, \qquad j = 1, \ldots, m$$

for any connected component Γ_j *of* Γ, *there exists a field, field* $E \in C^1(G^L) \cap C(\bar{G}^L)$, *such that*

$$\operatorname{curl} E = i\omega\mu^L H, \qquad \operatorname{div} E = 0 \quad \text{in } G^L,$$

$$E(x) = O\left(\frac{1}{|x|^\beta}\right), \quad \text{uniformly for } |x| \to \infty.$$

THEOREM 2. *Problem* (10) *has at most one solution.*

Proof. We consider (10) with homogeneous data $c = 0$, $g = 0$, $d = 0$. Since H^E is a solution of the vector Helmholtz equation with wave number k in G^E, the divergence of H^E solves the scalar Helmholtz equation with the same wave number k in G^E. From div $H^E|_\Gamma = d = 0$ and $\operatorname{Im}(k) > 0$ it follows that div H^E vanishes identically in G^E [1]. Using the identity curl curl = grad div $-\Delta$ and defining E^E by $E^E = 1/\sigma^E$ curl H^E, we conclude that H^E, E^E are solutions of the Maxwell equations in G^E.

On the other hand, H^L is a harmonic vector field in G^L, satisfying $H^L(x) = o(1)$ uniformly for $|x| \to \infty$. Therefore, we immediately get $H^L(x) = O\left(1/|x|^2\right)$ uniformly for $|x| \to \infty$. In addition to this, we deduce from the transmission conditions of (10)

$$\mu^L \int_{\Gamma_j} n \cdot H^L ds = \int_{\Gamma_j} n \cdot (\mu^L H^L) ds = \int_{\Gamma_j} n \cdot (\mu^E H^E) ds$$

$$= \frac{1}{i\omega} \int_{\Gamma_j} n \cdot (i\omega\mu^E H^E) ds = \frac{1}{i\omega} \int_{\Gamma_j} n \cdot \operatorname{curl} E^E ds = 0.$$

Applying Lemma 1 with $\beta = 1$ to H^L proves the existence of a field E^L defined in G^L, having the following properties:

$$\text{curl } E^L = i\omega\mu^L H^L \quad \text{in } G^L,$$

$$E^L(x) = o(1) \quad \text{uniformly for } |x| \to \infty.$$

Thus we have shown that H^L, E^L, H^E, E^E solve the homogeneous equations (3)–(5) with vanishing circulations

$$\int_{\gamma_i^L} \tau \cdot H^L dl = 0.$$

Now, from Theorem 1 we conclude $H^L \equiv 0$, $H^E \equiv 0$. □

Before we start with the existence proof for (10), we have to introduce some notation.

Definition. Let $0 < \alpha < 1$.

$$C^{0\alpha}(G), \qquad \|u\|_{0\alpha,G} = \sup_{x\in G} |u(x)| + \sup_{\substack{x\neq y\\ x,y\in G}} \frac{|u(x)-u(y)|}{|x-y|^\alpha},$$

is the space of Hölder-continuous functions on G.

$$V^{0\alpha}(\Gamma) = (C^{0\alpha}(\Gamma))^3, \qquad \|a\|_{V\alpha,\Gamma} = \max_{i=1,2,3}(\|a_i\|_{0\alpha,\Gamma}),$$

is the space of Hölder-continuous vector fields on Γ.

$$T^{0\alpha}(\Gamma) = \{a \in V^{0\alpha}(\Gamma) | n \cdot a = 0\}, \qquad \|a\|_{T\alpha,\Gamma} = \|a\|_{V\alpha,\Gamma},$$

is the space of Hölder-continuous tangential fields on Γ.

$$T_d^{0\alpha}(\Gamma) = \{a \in T^{0\alpha}(\Gamma) | \text{Div } a \in C^{0\alpha}(\Gamma)\}, \qquad \|u\|_{d\alpha,\Gamma} = \max(\|u\|_{T\alpha,\Gamma}, \|\text{Div } u\|_{0\alpha,\Gamma}),$$

is the space of Hölder-continuous tangential fields on Γ having Hölder-continuous surface divergence.

$$X_d^{0\alpha}(\Gamma) = T_d^{0\alpha}(\Gamma) \times C^{0\alpha}(\Gamma) \times C^{0\alpha}(\Gamma),$$

$$\|u\|_{X_{d\alpha}} = \max(\|u_1\|_{d\alpha,\Gamma}, \|u_2\|_{0\,\alpha\Gamma}, \|u_3\|_{0\alpha,\Gamma}),$$

$$\Phi(x,y) = \frac{1}{4\pi}\frac{e^{ik|x-y|}}{|x-y|}, \qquad \Phi_0(x,y) = \frac{1}{4\pi}\frac{1}{|x-y|}.$$

The spaces $C^{0\alpha}(G)$, $V^{0\alpha}(\Gamma)$, $T^{0\alpha}(\Gamma)$, $T_d^{0\alpha}(\Gamma)$, and $X_d^{0\alpha}$ equipped with the corresponding norms are Banach spaces.

Theorem 3. *For any $c \in T_d^{0\alpha}(\Gamma), g \in C^{0\alpha}(\Gamma), d \in C^{0\alpha}(\Gamma)$, problem* (10) *is uniquely solvable. The solution depends continuously on the given data.*

Proof. The proof will be divided into three parts. In the first part, we use a special ansatz for H^L and H^E to transform the transmission boundary-value problem into a boundary integral equation. In the second step, we show that the integral equation is of second kind. Finally we conclude the proof by showing the injectivity of the corresponding integral operator.

The ansatz. We are looking for solutions H^L, H^E of the form

$$\begin{aligned} & H^L(x) = \operatorname{grad}_x \int_\Gamma \lambda(y)\Phi_0(x,y)ds(y), \\ (11) \qquad & H^E(x) = \operatorname{curl}_x \int_\Gamma a(y)\Phi(x,y)ds(y) + \operatorname{grad}_x \int_\Gamma \lambda(y)\Phi(x,y)ds(y) \\ & \qquad + \int_\Gamma n(y)\delta(y)\Phi(x,y)ds(y), \end{aligned}$$

$a \in T_d^{0\alpha}(\Gamma)$, $\lambda \in C^{0\alpha}(\Gamma)$, $\delta \in C^{0\alpha}(\Gamma)$. For H^L, H^E we get [1] and [9]:

$$\text{(i)} \quad \begin{aligned} & H^L \in C^\infty(G^L) \cap C^{0\alpha}(\bar{G}^L), \qquad H^L(x) = O\left(\frac{1}{|x|^2}\right) \quad \text{uniformly for } |x| \to \infty, \\ & \int_{\gamma_i^L} \tau \cdot H^L dl = 0, \qquad i = 1, \dots, p. \end{aligned}$$

$$\text{(ii)} \quad H^E \in C^2(G^E) \cap C^{0\alpha}(\bar{G}^E), \quad \operatorname{div} H^E \in C^{0\alpha}(\bar{G}^E), \quad \operatorname{curl} H^E \in C^{0\alpha}(\bar{G}^E),$$

$$\text{(iii)} \quad \left.\begin{aligned} \operatorname{curl} H^L &= 0 \\ \operatorname{div} H^L &= 0 \end{aligned}\right. \text{ in } G^L, \qquad \left.\begin{aligned} (\Delta + k^2)H^E &= 0 \\ k^2 &= i\omega\sigma^E\mu^E \end{aligned}\right. \text{ in } G^E,$$

$$\text{(iv)} \quad \begin{aligned} & \|H^L\|_{V\alpha,\bar{G}^L} \le c_\alpha \|\lambda\|_{0\alpha,\Gamma}, \\ & \max(\|H^E\|_{V\alpha,\bar{G}^E}, \|\operatorname{div} H^E\|_{0\alpha,\bar{G}^E}) \le c_\alpha \left\| \begin{pmatrix} a \\ \lambda \\ \delta \end{pmatrix} \right\|_{X_{d\alpha}}, \\ & \|\operatorname{curl} H^E\|_{V\alpha,\bar{G}^E} \le c_\alpha \max(\|a\|_{d\alpha,\Gamma}, \|\delta\|_{0\alpha,\Gamma}). \end{aligned}$$

From (i)–(iv) we see that our ansatz meets all the regularity requirements of (10). H^L and H^E depend continuously on a, λ, and δ and solve the required differential equations. Therefore we only have to adjust the boundary values on Γ corresponding to (10).

Defining $F_\pm(x) = \lim_{h\searrow 0} F(x \pm hn(x))$, $x \in \Gamma$, n outer normal to G^E, we get the following jump conditions for both single and double layer potentials and their derivatives [1]:

$$\mathrm{curl}_x \int_\Gamma a(y)\Phi(x,y)ds(y)\Big|_\pm = \int_\Gamma \mathrm{curl}_x(a(y)\Phi(x,y))ds(y) \mp \frac{1}{2}n(x)\wedge a(x),$$

$$\mathrm{grad}_x \int_\Gamma \lambda(y)\Phi(x,y)ds(y)\Big|_\pm = \int_\Gamma \lambda(y)\mathrm{grad}_x\Phi(x,y)ds(y) \mp \frac{1}{2}n(x)\lambda(x),$$

$$\int_\Gamma n(y)\delta(y)\Phi(x,y)ds(y)\Big|_\pm = \int_\Gamma n(y)\delta(y)\Phi(x,y)ds(y),$$

$$\int_\Gamma \lambda(y)\Phi(x,y)ds(y)\Big|_\pm = \int_\Gamma \lambda(y)\Phi(x,y)ds(y),$$

$$\int_\Gamma \delta(y)\partial_{n_y}\Phi(x,y)ds(y)\Big|_\pm = \int_\Gamma \delta(y)\partial_{n_y}\Phi(x,y)ds(y) \pm \frac{1}{2}\delta(x).$$

The jump conditions do not change if we replace Φ by Φ_0 on both sides. Using

$$\begin{aligned}(\mathrm{div}\ H^E)(x) &= \Delta \int_\Gamma \lambda(y)\Phi(x,y)ds(y) + \int_\Gamma \mathrm{div}_x(n(y)\delta(y)\Phi(x,y))ds(y)\\ &= -k^2\int_\Gamma \lambda(y)\Phi(x,y)ds(y) + \int_\Gamma \delta(y)n(y)\cdot\mathrm{grad}_x\Phi(x,y)ds(y)\\ &= -k^2\int_\Gamma \lambda(y)\Phi(x,y)ds(y) - \int_\Gamma \delta(y)\partial_{n_y}\Phi(x,y)ds(y), \qquad x\in G^E,\end{aligned}$$

we deduce

$$H^L_+(x) = \int_\Gamma \lambda(y)\mathrm{grad}_x\Phi_0(x,y)ds(y) - \frac{1}{2}n(x)\lambda(x),$$

$$\begin{aligned}H^E_-(x) = &\int_\Gamma \mathrm{curl}_x(a(y)\Phi(x,y))ds(y) + \frac{1}{2}n(x)\wedge a(x)\\ &+\int_\Gamma \lambda(y)\mathrm{grad}_x\Phi(x,y)ds(y) + \frac{1}{2}n(x)\lambda(x)\\ &+\int_\Gamma n(y)\delta(y)\Phi(x,y)ds(y),\end{aligned}$$

$$\begin{aligned}(\mathrm{div}\ H^E)_-(x) = &-k^2\int_\Gamma \lambda(y)\Phi(x,y)ds(y)\\ &-\int_\Gamma \delta(y)\partial_{n_y}\Phi(x,y)ds(y) + \frac{1}{2}\delta(x).\end{aligned}$$

Introducing the operators

$$(Ma)(x) = 2n(x) \wedge \int_\Gamma \operatorname{curl}_x(a(y)\Phi(x,y))ds(y),$$

$$(Na)(x) = 2n(x) \cdot \int_\Gamma \operatorname{curl}_x(a(y)\Phi(x,y))ds(y),$$

$$(K\lambda)(x) = 2\int_\Gamma \lambda(y)\partial_{n_y}\Phi(x,y)ds(y),$$

$$(K'\lambda)(x) = 2\int_\Gamma \lambda(y)\partial_{n_x}\Phi(x,y)ds(y),$$

$$(S\lambda)(x) = 2\int_\Gamma \lambda(y)\Phi(x,y)ds(y),$$

$$(P\lambda)(x) = 2n(x) \wedge \int_\Gamma n(y)\lambda(y)\Phi(x,y)ds(y),$$

$$(Q\lambda)(x) = 2n(x) \cdot \int_\Gamma n(y)\lambda(y)\Phi(x,y)ds(y),$$

$$(R\lambda)(x) = 2n(x) \wedge \int_\Gamma \lambda(y)\operatorname{grad}_x\Phi(x,y)ds(y),$$

$$(K_0'\lambda)(x) = 2\int_\Gamma \lambda(y)\partial_{n_x}\Phi_0(x,y)ds(y),$$

$$(R_0\lambda)(x) = 2n(x) \wedge \int_\Gamma \lambda(y)\operatorname{grad}_x\Phi_0(x,y)ds(y),$$

we get the following expressions for the boundary values of H^L and H^E:

$$2n(x) \wedge H_+^L(x) = (R_0\lambda)(x),$$

$$2n(x) \wedge H_-^E(x) = (Ma)(x) - a(x) + (R\lambda)(x) + (P\delta)(x),$$

$$2n(x) \cdot (\mu^L H_+^L(x)) = \mu^L((K_0'\lambda)(x) - \lambda(x)),$$

$$2n(x) \cdot (\mu^E H_-^E(x)) = \mu^E((Na)(x) + (K'\lambda)(x) + \lambda(x) + (Q\delta)(x)),$$

$$2(\operatorname{div} H^E)_-(x) = -k^2(S\lambda)(x) + \delta(x) - (K\delta)(x).$$

Thus, we immediately see that solving (10) is equivalent to solving the following integral equation

$$Av = b, \quad A = \begin{pmatrix} M - I & R - R_0 & P \\ \mu^E N & \mu^E(I + K') + \mu^L(I - K_0') & \mu^E Q \\ 0 & -k^2 S & I - K \end{pmatrix},$$

(12)

$$v = \begin{pmatrix} a \\ \lambda \\ \delta \end{pmatrix}, \quad b = 2\begin{pmatrix} c \\ g \\ d \end{pmatrix}.$$

The integral equation (12) is of second kind in $X_d^{0\alpha}(\Gamma)$. According to [1]–[3] and [9] the operators defined above have the following mapping properties

$$M : T^{0\alpha}(\Gamma) \to T^{0\alpha}(\Gamma) \quad \text{respectively, } T_d^{0\alpha}(\Gamma), \qquad N : T_d^{0\alpha}(\Gamma) \to C^{0\alpha}(\Gamma),$$

$$Q, S, K, K', K_0' : C^{0\alpha}(\Gamma) \to C^{0\alpha}(\Gamma), \qquad P : C^{0\alpha}(\Gamma) \to T_d^{0\alpha}(\Gamma),$$

$$R, R_0 : C^{0\alpha}(\Gamma) \to T^{0\alpha}(\Gamma).$$

N, R, R_0 are continuous; $M, Q, S, K, K', K_0', P, R - R_0$ are compact.

Setting

$$F(x) = 2\int_\Gamma \lambda(y)\operatorname{grad}_x \Phi(x,y)ds(y), \qquad \lambda \in C^{0\alpha}(\Gamma), \quad x \in G^E,$$

we get

$$F \in C^2(G^E) \cap C^{0\alpha}(\bar{G}^E),$$

$$\operatorname{curl} F = 0 \quad \text{in } G^E,$$

$$n \wedge F|_\Gamma = R\lambda.$$

According to [1] we deduce

$$\operatorname{Div}(R\lambda) = \operatorname{Div}(n \wedge F) = -n \cdot \operatorname{curl} F|_\Gamma = 0$$

and therefore

$$\|R\lambda\|_{d\alpha,\Gamma} = \|R\lambda\|_{T\alpha,\Gamma}.$$

In the same way we show $\|R_0\lambda\|_{d\alpha,\Gamma} = \|R_0\lambda\|_{T\alpha,\Gamma}$, so that the continuity of R, R_0 and the compactness of $R - R_0$ carry over to the case where we consider R, R_0 as operators mapping $C^{0\alpha}(\Gamma)$ into $T_d^{0\alpha}(\Gamma)$.

Now A is split up into

$$A = B+C, \quad B = \begin{pmatrix} -I & 0 & 0 \\ \mu^E N & (\mu^E+\mu^L)I & 0 \\ 0 & 0 & I \end{pmatrix}, \quad C = \begin{pmatrix} M & R-R_0 & P \\ 0 & \mu^E K' - \mu^L K_0' & \mu^E Q \\ 0 & -k^2 S & -K \end{pmatrix}.$$

From the above considerations we immediately get

$$B, C : X_d^{0\alpha}(\Gamma) \to X_d^{0\alpha}(\Gamma),$$

where B is continuously invertible and C is compact. Therefore (12) is of second kind.

The operator A is injective. Consider a solution

$$v = \begin{pmatrix} a \\ \lambda \\ \delta \end{pmatrix} \in X_d^{0\alpha}(\Gamma)$$

of the homogeneous equation $Av = 0$. Inserting a, λ, δ in (11), the fields H^L and H^E obtained in this way solve the homogeneous problem (10). Corresponding to Theorem 2, they vanish identically. But H^L was defined as

$$H^L(x) = \operatorname{grad}_x \int_\Gamma \lambda(y)\Phi_0(x,y)ds(y)$$

and therefore

$$0 = 2n \cdot H_+^L = 2\int_\Gamma \lambda \partial_{n_x}\Phi_0 ds - \lambda = -(I - K_0')\lambda.$$

Since $G^L = \mathbb{R}^3 \backslash \bar{G}^E$ is connected and unbounded, we get $N(I - K_0') = \{0\}$ [4], so $\lambda = 0$. Using $\lambda = 0$ we obtain from the last component of $Av = b$, $(I - K)\delta = 0$. According to [1], $N(I - K) = \{0\}$ for $\operatorname{Im}(k) > 0$, and thus $\delta = 0$. From the first component of $Av = b$ we get $(I - M)a = 0$, and again $a = 0$ because $N(I - M) = \{0\}$ for $\operatorname{Im}(k) > 0$ [1].

From the above considerations we conclude that our auxiliary problem (10) is equivalent to the integral equation $Av = b$. Since $c \in T_d^{0\alpha}(\Gamma)$, $g \in C^{0\alpha}(\Gamma)$, $d \in C^{0\alpha}(\Gamma)$, the right-hand side b lies in $X_d^{0\alpha}(\Gamma)$. Now A is injective in $X_d^{0\alpha}(\Gamma)$ and therefore, according to the Riesz theory, continuously invertible in $X_d^{0\alpha}(\Gamma)$. So $Av = b$ is solvable for any $b \in X_d^{0\alpha}(\Gamma)$, with v depending continuously on b. If we use the components of v to define H^L, H^E using (11), we get a solution of (10). By

$$\|H^L\|_{V\alpha,\bar{G}^L} \leq c_\alpha \|\lambda\|_{0\alpha,\Gamma},$$

$$\max(\|H^E\|_{V\alpha,\bar{G}^E}, \|\text{div } H^E\|_{0\alpha,\bar{G}^E}) \leq c_\alpha \left\| \begin{pmatrix} a \\ \lambda \\ \delta \end{pmatrix} \right\|_{X_d\alpha},$$

$$\|\text{curl } H^E\|_{V\alpha,\bar{G}^E} \leq c_\alpha \max(\|a\|_{d\alpha,\Gamma}, \|\delta\|_{0\alpha,\Gamma}),$$

we get the continuous dependence of H^L and H^E on the data c, g, d. □

Up to now we have shown the unique solvability of the auxiliary problem (10) for arbitrary $c \in T_d^{0\alpha}(\Gamma)$, $g \in C^{0\alpha}(\Gamma)$, $d \in C^{0\alpha}(\Gamma)$. With the help of this result, we want to prove existence and uniqueness for (3)–(5) under the additional assumption (6) of prescribed circulations for H^L.

LEMMA 2. (i) *Let $G^J \subset G^L$ be defined as above , $\rho \in C^{0\alpha}(\mathbb{R}^3), \mathrm{supp}(\rho) \subset G^J$. Then*

$$u(x) = \int_{G^J} \rho(y)\Phi_0(x,y)dy \in C^2(\mathbb{R}^3),$$

$$u(x) = O\left(\frac{1}{|x|}\right) \qquad \text{grad } u(x) = O\left(\frac{1}{|x|^2}\right), \quad |x| \to \infty.$$

(ii) *Consider $J_e \in C^1(\mathbb{R}^3)$, div $J_e = 0$, supp$(J_e) \subset G^J$. There exist*

$$H^J \in C^1(\mathbb{R}^3), \qquad E^J \in C^1(G^L) \cap C(\bar{G}^L),$$

$$\text{curl } H^J = J_e, \quad \text{div } H^J = 0, \quad H^J(x) = O\left(\frac{1}{|x|^2}\right), \quad \textit{uniformly for } |x| \to \infty,$$

$$\text{curl } E^J = i\omega\mu^L H^J, \quad \text{div } E^J = 0, \quad E^J(x) = O\left(\frac{1}{|x|}\right), \quad \textit{uniformly for } |x| \to \infty.$$

(iii) *Corresponding to the Neumann fields in G^L we have $E_i^Z \in C^1(G^L) \cap C(\bar{G}^L)$,*

$$\text{curl } E_i^Z = i\omega\mu^L Z_i^L, \quad \text{div } E_i^Z = 0, \quad E_i^Z(x) = O\left(\frac{1}{|x|}\right), \quad \textit{uniformly for } |x| \to \infty.$$

Proof. The first part is an easy consequence of some well-known properties of the Newtonian potentials in $\mathbb{R}^3$ [5].

In (ii), we assume $J_e \in C^1(\mathbb{R}^3)$, supp$(J_e) \subset G^J$. Therefore $J_e \in C^{0\alpha}(\mathbb{R}^3)$ and from (i) we get

$$A = \int_{G^J} J_e(y)\Phi_0(x,y)dy \in C^2(\mathbb{R}^3).$$

Defining H^J as $H^J = \text{curl } A \in C^1(\mathbb{R}^3)$, we see that

$$\text{div } H^J = 0, \qquad \text{curl } H^J = \text{curl curl } A = (\text{grad div } - \Delta)A = J_e.$$

Corresponding to (i), the components of H^J behave as $O(1/|x|^2)$ uniformly for $|x| \to \infty$.

Since

$$\int_{\Gamma_j} n \cdot H^J ds = \int_{\Gamma_j} n \cdot \text{curl } A \, ds = 0, \qquad j = 1, \dots, m,$$

we may apply Lemma 1 to H^J (with $\beta = 1$) and get the existence of E^J.

For the last part, we remark that $Z_i^L \in C^1(G^L) \cap C(\bar{G}^L), i = 1, \dots, p$, are harmonic vector fields in G^L satisfying

$$n \cdot Z_i^L = 0 \quad \text{on } \Gamma, \quad Z_i^L = O\left(\frac{1}{|x|^2}\right), \quad \text{uniformly for } |x| \to \infty.$$

Using Lemma 1 again with $\beta = 1$, the proof is completed. □

With the help of Lemma 2, we obtain the main result of this paper.

THEOREM 4. *Consider $J_e \in C^1(\mathbb{R}^3), \operatorname{div} J_e = 0, \operatorname{supp}(J_e) \subset G^J, \bar{G}^J \subset G^L, G^J$ bounded. Under these assumptions, problem (3)–(6) possesses a solution H^L, E^L, H^E, E^E. H^L, H^E, E^E are uniquely determined.*

Proof. Consider H^J, which is given by Lemma 2. Define h_i^J as $h_i^J = \int_{\gamma_i^L} \tau \cdot H^J dl, i = 1, \ldots, p$, and

$$H^Z = \sum_{j=1}^{p} \left(h_j^L - h_j^J\right) Z_j^L,$$

$$c = n \wedge \left(H^J + H^Z\right)\big|_\Gamma,$$

$$g = n \cdot \left(\mu^L \left(H^J + H^Z\right)\right)|_\Gamma = n \cdot \left(\mu^L H^J\right)\big|_\Gamma.$$

From $H^J \in C^1(\mathbb{R}^3)$ we deduce $H^J|_\Gamma \in C^{0\alpha}(\Gamma)$. For the surface divergence of $n \wedge H^J$ on Γ we get, according to [1],

$$\operatorname{Div}\,(n \wedge H^J) = -n \cdot \operatorname{curl} H^J\,|_\Gamma = -n \cdot J_e|_\Gamma = 0$$

because $\operatorname{supp}(J_e) \subset G^J, \bar{G}^J \subset G^L$.

For the Neumman fields Z_i^L holds [6]

$$n \wedge Z_i^L\big|_\Gamma \in C^{0\alpha}(\Gamma), \qquad \operatorname{Div}\,(n \wedge Z_i^L) = -n \cdot \operatorname{curl} Z_i^L\big|_\Gamma = 0,$$

and therefore $c \in T_d^{0\alpha}(\Gamma)$. On the other hand, $H^J \in C^1(\mathbb{R}^3)$ implies $g \in C^{0\alpha}(\Gamma)$. By Theorem 3 there exist unique fields $\widetilde{H}^L$, $\widetilde{H}^E$,

$$\widetilde{H}^L \in C^1(G^L) \cap C(\bar{G}^L),$$

$$\widetilde{H}^E \in C^2(G^E) \cap C(\bar{G}^E), \quad \operatorname{div} \widetilde{H}^E \in C(\bar{G}^E), \quad \operatorname{curl} \widetilde{H}^E \in C(\bar{G}^E),$$

solving

$$\left.\begin{array}{l} \operatorname{curl} \widetilde{H}^L = 0 \\ \operatorname{div} \widetilde{H}^L = 0 \end{array}\right. \text{ in } G^L, \qquad \left.\begin{array}{c} (\Delta + k^2)\widetilde{H}^E = 0 \\ k^2 = i\omega\sigma^E\mu^E \end{array}\right. \text{ in } G^E,$$

$$\begin{array}{r} n \wedge \widetilde{H}^E - n \wedge \widetilde{H}^L = c \\ n \cdot (\mu^E \widetilde{H}^E) - n \cdot (\mu^L \widetilde{H}^L) = g \\ \operatorname{div} \widetilde{H}^E = 0 \end{array} \quad \text{on } \Gamma,$$

$$\int_{\gamma_i^L} \tau \cdot \widetilde{H}^L dl = 0, \qquad i = 1, \ldots, p,$$

$$\widetilde{H}(x) = o(1), \qquad \text{uniformly for } |x| \to \infty.$$

From div $\widetilde{H}^E = 0$ on Γ and $\mathrm{Im}(k) > 0$ we get, in the same way as in the proof of Theorem 2, that $\widetilde{H}^E$ and $\widetilde{E}^E = (1/\sigma^E)\mathrm{curl}\ \widetilde{H}^E$ solve the time-harmonic Maxwell equations in G^E with coefficients μ^E, σ^E, and ω.

$\widetilde{H}^L$ is harmonic in G^L. From the proof of Theorem 3, we know that for $\widetilde{H}^L$ even the stronger condition $\widetilde{H}^L(x) = O(\frac{1}{|x|^2})$ uniformly holds for $|x| \to \infty$. In addition we have

$$\begin{aligned} i\omega\mu^L \int_{\Gamma_j} n \cdot \widetilde{H}^L ds &= i\omega\mu^E \int_{\Gamma_j} n \cdot \widetilde{H}^E ds - i\omega \int_{\Gamma_j} g ds \\ &= \int_{\Gamma_j} n \cdot \mathrm{curl}\ \widetilde{E}^E ds - i\omega\mu^L \int_{\Gamma_j} n \cdot H^J ds \\ &= \int_{\Gamma_j} n \cdot \mathrm{curl}\ \widetilde{E}^E ds - i\omega\mu^L \int_{\Gamma_j} n \cdot \mathrm{curl}\ A\, ds, \qquad j = 1, \ldots, m, \end{aligned}$$

where A is defined in the proof of Lemma 2. Since Γ_j, $j = 1, \ldots, m$ are closed surfaces, we conclude by Stokes's theorem,

$$\int_{\Gamma_j} n \cdot \widetilde{H}^L ds = 0, \qquad j = 1, \ldots, m.$$

Now Lemma 1 guarantees the existence of $\widetilde{E}^L \in C^1(G^L) \cap C(\bar{G}^L)$ with

$$\mathrm{curl}\ \widetilde{E}^L = i\omega\mu^L \widetilde{H}^L \quad \text{in } G^L, \quad \widetilde{E}^L(x) = O\left(\frac{1}{|x|}\right), \quad \text{uniformly for } |x| \to \infty.$$

Summarizing the results obtained for $\widetilde{H}^L$, $\widetilde{E}^L$, $\widetilde{H}^E$, $\widetilde{E}^E$, we have

$$\widetilde{H}^L, \widetilde{E}^L \in C^1(G^L) \cap C(\bar{G}^L), \qquad \widetilde{H}^E, \widetilde{E}^E \in C^1(G^E) \cap C(\bar{G}^E),$$

$$\left.\begin{aligned} \mathrm{curl}\ \widetilde{H}^L &= 0 \\ \mathrm{curl}\ \widetilde{E}^L &= i\omega\mu^L \widetilde{H}^L \end{aligned}\right\} \text{in } G^L, \qquad \left.\begin{aligned} \mathrm{curl}\ \widetilde{H}^E &= \sigma^E \widetilde{E}^E \\ \mathrm{curl}\ \widetilde{E}^E &= i\omega\mu^E \widetilde{H}^E \end{aligned}\right\} \text{in } G^E,$$

$$\left.\begin{aligned} n \wedge \widetilde{H}^E - n \wedge \widetilde{H}^L &= c \\ n \cdot (\mu^E \widetilde{H}^E) - n \cdot (\mu^L \widetilde{H}^L) &= g \end{aligned}\right\} \text{on } \Gamma,$$

$$\int_{\gamma_i^L} \tau \cdot \widetilde{H}^L dl = 0, \qquad i = 1, \ldots, p,$$

$$\widetilde{H}^L(x) = o(1), \quad \widetilde{E}^L(x) = o(1) \quad \text{uniformly for } |x| \to \infty.$$

But in Lemma 2 the existence of E^J, $E^Z \in C^1(G^L) \cap C(\bar{G}^L)$,

$$\mathrm{curl}\ E^J = i\omega\mu^L H^J, \qquad \mathrm{curl}\ E^Z = i\omega\mu^L H^Z,$$

is shown, both behaving uniformly like $o(1)$ for $|x| \to \infty$.

Defining H^L, E^L, H^E, E^E as

$$H^L = \widetilde{H}^L + H^J + H^Z, \quad E^L = \widetilde{E}^L + E^J + E^Z, \quad H^E = \widetilde{H}^E, \quad E^E = \widetilde{E}^E,$$

and using curl $H^J = J_e$, curl $H^Z = 0$, we get

$$\left.\begin{aligned} \operatorname{curl} H^L &= \operatorname{curl}\,(\widetilde{H}^L + H^J + H^Z) = J_e \\ \operatorname{curl} E^L &= \operatorname{curl}\,(\widetilde{E}^L + E^J + E^Z) = i\omega\mu^L H^L \end{aligned}\right. \quad \text{in } G^L,$$

$$\left.\begin{aligned} \operatorname{curl} H^E &= \sigma^E E^E \\ \operatorname{curl} E^E &= i\omega\mu^E H^E \end{aligned}\right. \quad \text{in } G^E,$$

$$\begin{aligned} n \wedge H^E &= n \wedge \widetilde{H}^E = n \wedge \widetilde{H}^L + c = n \wedge (\widetilde{H}^L + H^J + H^Z) = n \wedge H^L, \\ n \cdot (\mu^E H^E) &= n \cdot (\mu^E \widetilde{H}^E) = n \cdot (\mu^L \widetilde{H}^L) + g = n \cdot (\mu^L(\widetilde{H}^L + H^J + H^Z)) \quad \text{on } \Gamma, \\ &= n \cdot (\mu^L H^L), \end{aligned}$$

and

$$\int_{\gamma_i^L} \tau \cdot H^L dl = \int_{\gamma_i^L} \tau \cdot (\widetilde{H}^L + H^J + H^Z) dl = \int_{\gamma_i^L} \tau \cdot H^J dl + h_i^L - h_i^J = h_i^L, \qquad i = 1, \ldots, p.$$

Therefore, H^L, E^L, H^E, E^E solve (3)–(6). □

COROLLARY. For $J_e \in C^1(\mathbb{R}^3)$, div $J_e = 0$, supp$(J_e) \subset G^J$, $\bar{G}^J \subset G^L$ bounded, problem (3)–(5) is solvable.

In the homogeneous case $J_e = 0$ we get exactly p linear independent solutions H^L, H^E, E^E, where p denotes the topological genus of G^E, respectively, G^L.

E^L is not uniquely determined.

Proof. The first statement follows immediately from the last theorem by choosing the circulations h_i^L, $i = 1, \ldots, p$ arbitrarily.

In the case $J_e = 0$ Theorem 3 shows the existence of p solutions H_j^L, E_j^L, H_j^E, E_j^E, $j = 1, \ldots, p$, of (3), (4), (5), having circulations $h_{ji}^L = \delta_{ij}, i = 1, \ldots, p$. The linear independence of H_j^L, E_j^L, H_j^E is a consequence of the uniqueness results of Theorem 1.

The nonuniqueness of E^L is obvious. □

REFERENCES

[1] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, 1983.

[2] W. Knauff and R. Kress, *On the exterior boundary-value problem for the time-harmonic Maxwell equations*, J. Math. Anal. Appl., 72 (1979), pp. 215–235.

[3] R. Kress, *On the boundary operator in electromagnetic scattering*, Proc. Royal Society Edinburgh, Sect. A, 103 A (1986) pp. 91–98.

[4] ———, *Linear Integral Equations*, Springer-Verlag, New York, 1989.

[5] J. L. Lions and R. Dautray, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 1, Springer-Verlag, New York, 1990.

[6] E. Martensen, *Potentialtheorie*, Teubner, Leipzig, 1968.

[7] C. Müller, *Grundprobleme der mathematischen Theorie elektromagnetischer Schwingungen*, Springer-Verlag, New York, 1957.

[8] P. Werner, *Über das Verhalten elektromagnetischer Felder für kleine Frequenzen in mehrfach zusammenhängenden Gebieten*, J. Reine Angew. Math., 1 (1975), pp. 365–397 and 2 (1976), pp. 98–121.

[9] P. Wilde, *Über Transmissionsprobleme bei der vektoriellen Helmholtzgleichung*, Ph.D. thesis, Univ. of Göttingen, Germany, 1985.

SIAM J. MATH. ANAL.
Vol. 24, No. 6, pp. 1458–1514, November 1993

004

DETERMINATION OF SCATTERING FREQUENCIES FOR AN ELASTIC FLOATING BODY*

C. HAZARD† AND M. LENOIR†

Abstract. This paper studies the time-harmonic motions of a three-dimensional elastic floating body on the sea, in the case of finite and constant depth. In order to compute the resonant states of such a system, a variational formulation for the determination of the scattering frequencies of the problem is investigated, i.e., the poles of the analytic continuation of the solution operator. A practical method, based on a series expansion of the solution in a vicinity of infinity, is described. The scattering frequencies are shown to be the solutions of a nonlinear eigenvalue problem for a compact operator. Numerical results for a two-dimensional model are presented.

Key words. fluid-structure interaction, scattering frequencies, resonances, series solutions

AMS subject classifications. 35B60, 35C10, 35P25, 47A55, 76B15

1. Introduction.

1.1. Motivation. In a previous paper [18], we described two practical methods for the computation of the scattering frequencies of the Helmholtz equation in exterior domains, i.e., the poles of the analytical continuation of the resolvent of the problem. In both cases, we constructed explicitly this analytical continuation by reducing the problem to an equivalent one, set in a bounded domain, using either an integral representation or a series expansion of the solution in the vicinity of infinity. We have shown that the determination of the scattering frequencies then amounts to the solution of a nonlinear and nonselfadjoint eigenvalue problem for a compact operator.

The purpose of the present paper is to show how one of these methods applies to the problem of the linearized motions of an elastic floating body on the sea, in the case of finite and constant depth. The determination of the scattering frequencies, as well as the associated scattering modes, provides a new and especially convenient method for the study of the dynamic stability of ships, in the context of linearized approximation. Indeed, instead of computing the response of the system for numerous cases of external forces (which essentially depend on the frequency and the direction of propagation of the incident wave), our method consists in determining intrinsic quantities of the system : the scattering frequencies, which are characteristic values of the problem. As described in [18], their knowledge allows not only to locate the peaks of the response curve of the floating body, i.e., the (real) frequencies of the incident waves for which the energy transmitted to the body is maximum, but also to obtain an a priori estimate of the response of the system in the vicinity of such frequencies.

Our analysis is based on a stationary approach of scattering frequencies, since the starting point of our study consists in the time-harmonic vibration problem: the scattering frequencies are shown to be the poles of the analytic continuation for complex frequencies of the solution operator associated with this problem. An equivalent approach (see [2], [5], [34]), which is closely related to the time-dependent scattering theory of Lax and Phillips [15], offers another characterization of scattering frequencies:

*Received by the editors December 7, 1992; accepted for publication March 8, 1993.

†Laboratoire de Simulation et Modélisation des Phénomènes de Propagation (associé au Centre National de la Recherche Scientifique et à l'Université Pierre et Marie Curie), Ecole Nationale Supérieure de Techniques Avancées, Chemin de la Hunière, 91120 Palaiseau, France.

they appear as the poles of the analytic continuation of the so-called scattering matrix, which connects the asymptotic behaviours of the incident and scattered waves. The agreement of both approaches is proved in [5] in the case of a rigid floating body. As we will see, the stationary approach is particularly well adapted for a numerical treatment of the problem.

The characterization of scattering frequencies which is proposed in this paper proceeds from the coupling method between variational formulation and series expansion (which leads to the so-called "localized finite element" method; see Richer [28]). It is in fact a generalization to the three-dimensional case and to complex frequencies of the method described by Lenoir and Tounsi [17] (in the two-dimensional case, the extension to complex frequencies is studied in [5]). An alternative approach, which applies in both finite and infinite depth cases, will be presented in [6]: it is based on the coupling method between variational formulation and integral representation introduced by Jami and Lenoir [9] for the sea-keeping problem.

For other fluid structure interaction problems, some asymptotic properties of scattering frequencies are studied by Ohayon and Sanchez-Palencia [23], Sanchez-Hubert and Sanchez-Palencia [29], [30], and Vullierme-Ledard [32]. Also worth mentioning is the work of Joly and Poisson [12], [25], [26], who deal with the scattering of acoustic and elastic waves: they compute the scattering frequencies by means of an integral equation method.

1.2. Description of the method. The general organization of the paper is as follows.

In §2, we introduce the linearized equations describing the small motions of an elastic body floating on the sea and subject to an incident monochromatic wave. These equations involve the velocity potential φ of the scattered wave which is defined in the unbounded fluid domain Ω, and the displacement vector field u of the body. The "stationary problem" then consists of determining the pair (φ, u) for a given frequency (whose square is denoted by ν) and a given incident wave. The purpose of the present paper is to construct explicitly the analytic extension to complex ν of the "solution operator" $\mathcal{R}_\nu$ associated with this problem, i.e., the linear operator which maps the incident wave onto the solution (φ, u). The scattering frequencies actually are the poles of the analytic continuation of $\mathcal{R}_\nu$ in the complex plane $\mathbb{C}$.

In a first step (§2.4), we show how $\mathcal{R}_\nu$ can be easily extended to the upper complex half plane $\mathbb{C}^+ = \{\nu \in \mathbb{C};\ \operatorname{Im} \nu > 0\}$: in fact, the stationary problem extends in this case to a well-posed coercive problem. In order to construct the continuation of $\mathcal{R}_\nu$ in the lower complex half plane $\mathbb{C}^-$, we proceed as in [18] by reducing the initial problem to an equivalent one, which is set in a bounded fluid domain $\hat{\Omega}$ delimited by a fictitious, vertical, cylindrical boundary Σ (§2.5). This reduction is performed by noticing that the only knowledge of the restriction $\varphi_{|\Sigma}$ of φ allows us to determine φ in the whole domain $\check{\Omega}$ outside Σ. We can then define a "reduced problem" by means of a "coupling condition" on Σ which implies the analytical matching between $\varphi_{|\hat{\Omega}}$ and $\varphi_{|\check{\Omega}}$.

The extension of the coupling condition to complex values of ν, which is dealt with in §3, is the backbone of the analytic continuation of $\mathcal{R}_\nu$. To construct this extension, we exhibit an explicit form of the coupling condition by means of a series expansion which follows from the particular choice of the outer domain $\hat{\Omega}$. Indeed, this choice enables us to expand $\varphi_{|\check{\Omega}}$ as a series of functions with separated variables. In the case of real positive ν (§3.1), the application of the method of separation of variables is rather classical: the determination of the solutions with separated variables amounts to solving the so-called dispersion equation $-\zeta \tan \zeta = \nu$, and the completeness of these solutions results from standard arguments of spectral theory. The properties of the series expansion of $\varphi_{|\tilde{\Omega}}$ in terms of solutions with separated variables are studied in Appendix A. The generalization of this method for complex ν (§3.2) causes two difficulties. The first one lies in the solution of the dispersion equation. We make use of an original method for solving some transcendental equations (Henrici [7]), which is based on the properties of Cauchy integrals on arcs. It provides an explicit form of the solutions which actually are branches of an algebraic function having a countable infinity of singularities of order 1. For the sake of clarity, the analysis of the dispersion equation is postponed to Appendix B. The second difficulty is the completeness of the solutions with separated variables, which does not follow any more from classical spectral theory, since it involves the completeness of a nonorthogonal basis. Following Kato [13], we use perturbation techniques for orthonormal families. Finally, we prove in §3.3 that the coupling condition depends analytically on ν in a subset of the complex plane.

This latter property allows us to proceed to the analytic continuation of the reduced problem (§4) by writing it as a Fredholm equation, which requires a sharp study of the series expansion of the coupling condition. Using a theorem due to Steinberg [31], we deduce that its solution depends meromorphically on ν : the poles, which are located in $\mathbb{R}^+$ or $\mathbb{C}^-$, are the solutions of a nonlinear eigenvalue problem. By virtue of the equivalence between the initial and reduced problems, we finally prove that these poles are nothing but the poles of the analytic continuation of the solution operator $\mathcal{R}_\nu$: these are the eigenfrequencies (in $\mathbb{R}^+$) and the scattering frequencies (in $\mathbb{C}^-$) of the problem.

Finally, we present in §5 the outcome of a numerical application for a two-dimensional "catamaran,"i.e., two rigid hulls linked together by an elastic beam. We briefly describe the main stages of the method in this case. Numerical results are shown: they illustrate in particular the connection between the scattering frequencies and the peaks of the response curve of the system.

2. Linearized equations of the coupled problem.

2.1. Notation. Consider an elastic body which floats (without forward motion) on the free surface of an inviscid, incompressible fluid. The motion of the fluid is assumed irrotational. When the system is at rest, the fluid fills an unbounded domain $\Omega \subset \mathbb{R}^3$ whose boundary $\partial\Omega$ consists of the free surface S, the bottom F, which is supposed to be plane and parallel to S, and the immersed surface Γ of the body (see Fig. 2.1). The body fills a bounded connected domain $B \subset \mathbb{R}^3$ with Lipschitz boundary ∂B; the emerged part of ∂B is denoted by Γ_0: it is assumed free. The coordinates (x_1, x_2, x_3) of every point x of the fluid or the body are expressed in an orthonormal system $(O, \vec{x}_1, \vec{x}_2, \vec{x}_3)$ chosen such that $(O, \vec{x}_1, \vec{x}_2)$ contains the free surface S and $(O, \vec{x}_3)$ is the ascending vertical axis. We denote by n either the outer unitary normal to $\partial\Omega$ or the inner normal to ∂B (which coincide on Γ).

All quantities involved in the problem are supposed to be dimensionless. Indeed, we can rescale all physical quantities by means of three independent characteristic constants: the fluid density (which is assumed constant in Ω), the gravitational constant, and the depth. In particular, the equation of the bottom F is $x_3 = -1$.

In the sequel, we will use the following notation. For any open set $D \subset \mathbb{R}^3$, the usual scalar product and the associated norm in $L^2(D)$ (respectively, the Sobolev space $H^s(D)$, for $s \in \mathbb{R}$) are denoted by $(\cdot\,,\cdot)_D$ and $\|\cdot\|_D$ (respectively, $(\cdot\,,\cdot)_{s,D}$ and $\|\cdot\|_{s,D}$). For an unbounded domain Ω, $H^1_{\text{loc}}(\Omega)$ denotes the Frechet space of functions φ such that $\varphi_{|D} \in H^1(D)$ for any bounded domain $D \subset \Omega$. If Σ is the boundary of

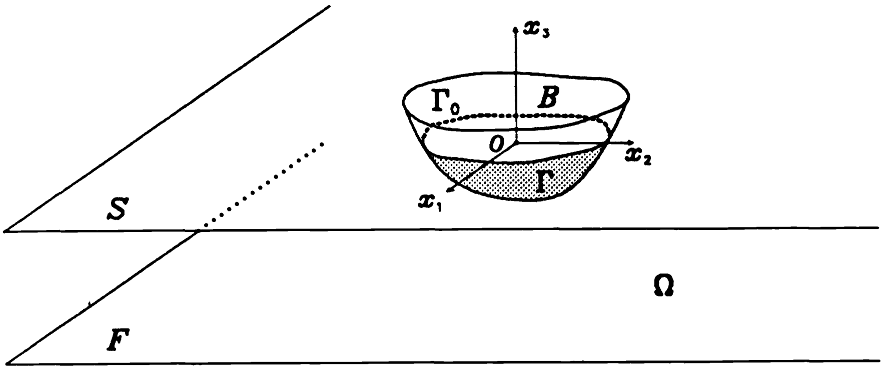


FIG. 2.1

an open set $D \subset \mathbb{R}^3$, the semiduality product between $H^{-s}(\Sigma)$ and $H^s(\Sigma)$ for $s > 0$ is denoted by $\langle \cdot , \cdot \rangle_{s,\Sigma}$.

2.2. The time-dependent problem. In this paper, we are concerned with the linearized approximation of the problem: the amplitudes of the motions of the fluid and the body are assumed small with respect to the dimensions of the body. Let $\Phi(x,t)$ be the velocity potential of the fluid and $U(x,t)$ the displacement field of the body. Let $e_{ij}(U)$ and $\sigma_{ij}(U)$ (for $i,j = 1,3$) be, respectively, the components of the strain and stress tensors given by

$$e_{ij}(U) = \tfrac{1}{2}\left(\partial_{x_j} U_i + \partial_{x_i} U_j\right) \quad \text{and} \quad \sigma_{ij}(U) = a_{ijkh} e_{kh}(U),$$

where $a_{ijkh} \in L^\infty(B)$ are the elastic coefficients. Note that we make use of the classical summation convention for twice-repeated indices in a product. In the case of an isotropic material, we have

$$a_{ijkh} = \lambda\, \delta_{ij}\delta_{kh} + \mu\,(\delta_{ih}\delta_{jk} + \delta_{ik}\delta_{jh}),$$

where λ and μ are the so-called Lamé coefficients. In the sequel, we will only assume that the elastic coefficients satisfy the symmetry and positivity properties

$$a_{ijkh} = a_{jikh} = a_{khij}, \tag{2.1}$$

$$a_{ijkh}(x)\, e_{ij} e_{kh} \geq \alpha\, e_{ij} e_{ij} \quad \forall x \in B \tag{2.2}$$

for any real symmetric tensor e_{ij}, where α is a positive constant. This amounts to saying that $2\mu(x) \geq \alpha$ and $3\lambda(x) + 2\mu(x) > \alpha$ for an isotropic material.

The linearized equations of the time-dependent problem are the following (a mathematical formulation of this problem in terms of semigroups of linear operators, as well as an existence and uniqueness result, are given by Licht [19]) :

$$\Delta\Phi = 0 \quad \text{in } \Omega, \tag{2.3}$$

$$\partial_t^2 \Phi + \partial_n \Phi = 0 \quad \text{on } S, \tag{2.4}$$

$$\partial_n \Phi = 0 \quad \text{on } F, \tag{2.5}$$

$$\rho\, \partial_t^2 U_i - \partial_{x_j} \sigma_{ij}(U) = 0 \quad \text{in } B, \tag{2.6}$$

$$\sigma_{ij}(U)\, n_j = 0 \quad \text{on } \Gamma_0, \tag{2.7}$$

$$\partial_t U.n - \partial_n \Phi = 0 \quad \text{on } \Gamma, \tag{2.8}$$

$$\sigma_{ij}(U)\, n_j - \partial_t \Phi\, n_i = 0 \quad \text{on } \Gamma. \tag{2.9}$$

Laplace equation (2.3) follows from the mass conservation law; the free surface condition (2.4) combines both the kinematic and dynamic relations to be satisfied on S (John [10]); (2.5) expresses that the normal velocity vanishes on the bottom F. Relations (2.6) are the classical dynamic equations in a continuous medium (see, e.g., Nečas and Hlaváček [22]), where $\rho \in L^\infty(B)$ is the (dimensionless) mass density of the body, which is assumed to satisfy

$$\rho(x) \geq \rho_0 \quad \text{in } B \tag{2.10}$$

for some positive constant ρ_0. Equations (2.7) express that the emerged part of ∂B is free. Finally, (2.8) and (2.9) are, respectively, the kinematic and dynamic coupling conditions on Γ (i.e., the continuity of the normal velocity and of the pressure).

In the case of a rigid body (i.e., $e_{ij}(U) = 0$ for $i, j = 1, 3$), the linearized displacement field reads $U = a + b \times x$ (see [22]): the system (2.3)–(2.9) then reduces to the problem studied by John [10].

2.3. The stationary problem. We consider the time-harmonic vibration of the system when it is subject to a monochromatic incident wave of pulsation $\omega > 0$. We are thus led to seek a solution of (2.3) to (2.9) in the form

$$\Phi(x,t) = \mathrm{Re}\left((\varphi(x) + \varphi_I(x))e^{-i\omega t}\right) \quad \text{and} \quad U(x,t) = \mathrm{Re}\left(i\, u(x)e^{-i\omega t}\right), \tag{2.11}$$

where $\varphi_I(x)$ is the velocity potential of the incident wave, that is, a solution of

$$\begin{aligned} &\Delta\varphi_I = 0 \quad \text{in } \{(x_1,x_2,x_3) \in \mathbb{R}^3;\ -1 < x_3 < 0\}, \\ &\partial_n\varphi_I - \nu\,\varphi_I = 0 \quad \text{on } x_3 = 0, \\ &\partial_n\varphi_I = 0 \quad \text{on } x_3 = -1, \end{aligned}$$

where $\nu = \omega^2$. For instance, in the case of a plane wave which propagates in the direction k (unit vector of the plane $(\vec{x}_1, \vec{x}_2)$), we have $\varphi_I(x) = A\cosh(\nu_0(x_3+1))\, e^{i\nu_0 x\cdot k}$, where A is a complex constant and ν_0 is the only positive root of equation

$$\nu_0 \tanh \nu_0 = \nu. \tag{2.12}$$

Substituting the expressions (2.11) of Φ and U in (2.3) to (2.9), we obtain

$$\begin{aligned} &\Delta\varphi = 0 \quad \text{in } \Omega, \\ &\partial_n\varphi - \nu\,\varphi = 0 \quad \text{on } S, \\ &\partial_n\varphi = 0 \quad \text{on } F, \\ &\nu\rho\, u_i + \partial_{x_j}\sigma_{ij}(u) = 0 \quad \text{in } B, \\ &\sigma_{ij}(u)\, n_j = 0 \quad \text{on } \Gamma_0, \\ &\nu^{1/2}\, u.n - \partial_n\varphi = f' \quad \text{on } \Gamma, \\ &\sigma_{ij}(u)\, n_j + \nu^{1/2}\,\varphi\, n_i = n_i f'' \quad \text{on } \Gamma, \end{aligned} \tag{2.13}$$

where $f' = \partial_n\varphi_I$ and $f'' = -\nu^{1/2}\varphi_I$ are given functions on Γ. Note that the complex factor i in the definition (2.11) of u avoids complex coefficients in the boundary conditions on Γ.

In addition, we must specify the asymptotic behaviour of φ by means of the outgoing radiation condition which expresses that the energy (associated with the scattered wave) radiates towards infinity (see [11])

$$\lim_{R\to+\infty} \int_{\Sigma_R} \left|\partial_n\varphi - i\nu_0\,\varphi\right|^2 d\Sigma = 0, \quad \text{with } \Sigma_R = \{x \in \Omega;\ (x_1^2 + x_2^2) = R^2\}, \tag{2.14}$$

where ν_0 still is the only positive root of (2.12). Setting

$$\mathcal{H}_{\mathrm{loc}} = H^1_{\mathrm{loc}}(\Omega) \times H^1(B)^3 \quad \text{and} \quad \mathcal{F} = L^2(\Gamma)^2, \tag{2.15}$$

we thus define the "stationary problem,"denoted by $\mathcal{P}_\nu$ in the sequel:

$$\begin{array}{l} \text{Find } X = (\varphi, u) \in \mathcal{H}_{\mathrm{loc}} \text{ such that} \\ \varphi \text{ and } u \text{ satisfy (2.13) and} \\ \varphi \text{ satisfies the radiation condition (2.14),} \end{array} \tag{2.16}$$

where the datum $f = (f', f'')$ is assumed to belong to $\mathcal{F}$. We will see that $\mathcal{P}_\nu$ has a unique solution, except maybe for a discrete set of $\nu \in \mathbb{R}^+$ (the set of eigenvalues of the problem).

Remark 2.1. If ν is not an eigenvalue of the problem, the solution $X = (\varphi, u)$ of $\mathcal{P}_\nu$ actually provides the asymptotic behaviour when $t \to +\infty$ of the solution $(\Phi(t), U(t))$ of the time-dependent problem (2.3)–(2.9). More precisely, the quantities $\Phi(x,t) - \mathrm{Re}\,((\varphi(x) + \varphi_I(x))e^{-i\omega t})$ and $U(x,t) - \mathrm{Re}\,(i\,u(x)e^{-i\omega t})$ tend to 0: this result, which is called the "limiting amplitude principle," has been proved by Vullierme-Ledard [33] in the case of a fixed rigid body.

In the sequel, we denote by $\mathcal{R}_\nu$ the "solution operator" associated with problem $\mathcal{P}_\nu$, i.e., the linear operator which maps the datum $f \in \mathcal{F}$ onto the solution $X \in \mathcal{H}_{\mathrm{loc}}$ of $\mathcal{P}_\nu$. The aim of the present paper is to construct explicitly the extension of $\mathcal{R}_\nu$ to complex values of ν. We will show that it extends to a meromorphic function of ν whose poles are nothing but the scattering frequencies of the problem, and we will give a characterization of these poles which is well adapted for their numerical computation.

The extension of $\mathcal{R}_\nu$ is performed in two steps. The first one, which consists in extending $\mathcal{P}_\nu$ to the upper complex half plane (i.e., for $\mathrm{Im}\,\nu > 0$) does not raise any difficulty. It is dealt with in §2.4 below. On the other hand, the extension to the lower complex half plane requires a precise control of the asymptotic behaviour of φ when $|x| \to +\infty$ (which becomes exponentially increasing in this case). Following the same idea as in [18], we will reduce problem $\mathcal{P}_\nu$ to an equivalent problem set in a bounded domain (§2.5) in order to carry out this extension (§§3 and 4).

2.4. The solution operator. Let $\nu \in \mathbb{C}^+ = \{\nu \in \mathbb{C};\ \mathrm{Im}\,\nu > 0\}$ and let $(\cdot)^{1/2}$ denote the principal value of the complex square root, i.e.,

$$\nu^{1/2} = r^{1/2}\, e^{i\theta/2} \ \text{ if } \ \nu = r\, e^{i\theta} \quad \text{with } r > 0 \quad \text{and} \quad |\theta| < \pi.$$

Problem $\mathcal{P}_\nu$, defined in (2.16) for real positive ν can be extended to complex values $\nu \in \mathbb{C}^+$ as follows. Let $\mathcal{H}$ denote the Hilbert space

$$\mathcal{H} = H^1(\Omega) \times H^1(B)^3. \tag{2.17}$$

Consider then the problem

$$\begin{array}{l} \text{Find } X = (\varphi, u) \in \mathcal{H} \text{ such that} \\ \varphi \text{ and } u \text{ satisfy (2.13),} \end{array} \tag{2.18}$$

which amounts to replacing the radiation condition by a decay condition at infinity. This problem will also be denoted by $\mathcal{P}_\nu$ in the sequel.

PROPOSITION 2.1. *Let $\nu \in \mathbb{C}^+$. For every $f \in \mathcal{F}$, problem $\mathcal{P}_\nu$ has a unique solution $X = \mathcal{R}_\nu f$ in $\mathcal{H}$. Furthermore, there exists $K(\nu) > 0$ such that*

$$\|\mathcal{R}_\nu f\|_{\mathcal{H}} \le K(\nu) \|f\|_{\mathcal{F}} \quad \forall f \in \mathcal{F}. \tag{2.19}$$

Proof. It is an easy matter to prove that the following variational formulation is equivalent to $\mathcal{P}_\nu$:

$$\text{(2.20)} \qquad \begin{array}{l} \text{Find } X \in \mathcal{H} \text{ such that} \\ a_\nu(X,Y) = l(f;Y) \quad \forall Y \in \mathcal{H}, \end{array}$$

where $a_\nu(\cdot\,,\cdot)$ is the sesquilinear form defined on $\mathcal{H} \times \mathcal{H}$ by

$$a_\nu(X,Y) = k(X,Y) - \nu\, m(X,Y) - \nu^{1/2}\, c(X,Y),$$

with

$$\begin{aligned} k(X,Y) &= \int_\Omega \nabla\varphi.\nabla\bar{\psi}\, d\Omega + \int_B a_{ijkh} e_{ij}(u) e_{kh}(\bar{v})\, dB, \\ m(X,Y) &= \int_S \varphi\bar{\psi}\, dS + \int_B \rho\, u.\bar{v}\, dB, \\ c(X,Y) &= \int_\Gamma (u.n\,\bar{\psi} + \varphi\,\bar{v}.n)\, d\Gamma \end{aligned}$$

for all $X = (\varphi, u)$ and $Y = (\psi, v)$ in $\mathcal{H}$, and $l(f;\cdot)$ is the semilinear form

$$l(f;Y) = -\int_\Gamma (f'\bar{\psi} + f''\bar{v}.n)\, d\Gamma.$$

a_ν and l are of course continuous on $\mathcal{H}$. To see that a_ν is a coercive form, notice that

$$\text{(2.21)} \qquad \operatorname{Im}\left\{\nu^{-1/2}\, a_\nu(X,X)\right\} = -\operatorname{Im}(\nu^{1/2})\,\left(|\nu|^{-1} k(X,X) + m(X,X)\right),$$

from which we deduce

$$\text{(2.22)} \qquad |a_\nu(X,X)| \geq \left|\operatorname{Im}(\nu^{1/2})\right| \operatorname{Min}\left\{|\nu|^{-1/2}, |\nu|^{1/2}\right\} (k(X,X) + m(X,X)).$$

Moreover, $(k(X,X) + m(X,X))^{1/2}$ is a norm in $\mathcal{H}$ equivalent to $\|X\|_{\mathcal{H}}$. This follows on one hand from the assumptions (2.1), (2.2), and (2.10), and from Korn's inequality (see, e.g., Nečas and Hlaváček [22]):

$$\text{(2.23)} \qquad \int_B e_{ij}(u) e_{ij}(\bar{u})\, dB + \int_B |u|^2\, dB \geq \gamma\, \|u\|_{1,B}^2 \quad \forall\, u \in H^1(B).$$

On the other hand, we have

$$\text{(2.24)} \qquad \|\varphi\|_{1,\Omega}^2 \leq \|\nabla\varphi\|_\Omega^2 + \|\varphi\|_S^2 \quad \forall\, \varphi \in H^1(\Omega),$$

which can be proved by the same techniques as for Poincaré's inequality in a strip. The existence and uniqueness of the solution of (2.20) thus results from Lax–Milgram's theorem. Finally, inequality (2.19) is a simple consequence of the continuity of l and of the coerciveness of a_ν. □

Remark 2.2. This new problem actually defines a continuous extension in $\mathbb{C}^+$ of $\mathcal{P}_\nu$ initially constructed for real positive ν : when $\nu \in \mathbb{C}^+ \to \mu \in \mathbb{R}^+$, its solution tends (in $\mathcal{H}_{\text{loc}}$) to the solution of $\mathcal{P}_\mu$. This result, which is called the "limiting absorption principle," may be proved by classical techniques using continuity properties of the Green function with respect to ν (see [5] or [16] for the infinite depth case). In fact, it appears as a straightforward consequence of the analyticity results which will be

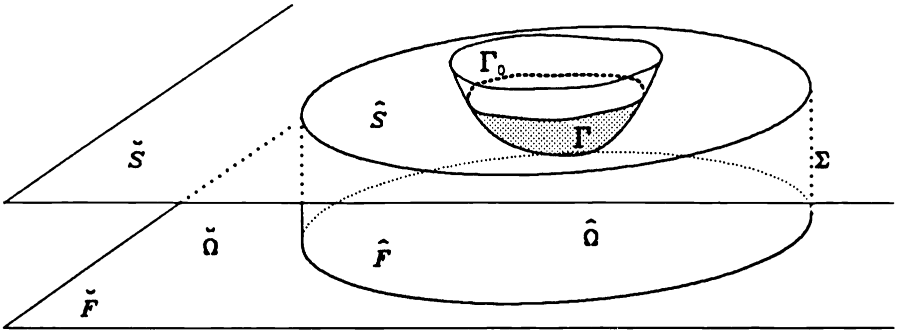


FIG. 2.2. *The inner and outer domains.*

worked out in §4: these results provide far more precise information about the ν-dependence of the solution of $\mathcal{P}_\nu$ than a simple continuity property.

Remark 2.3. If $\nu \in \mathbb{C}^- = \{\nu \in \mathbb{C};\ \operatorname{Im}\nu < 0\}$, problem (2.18) still is well posed: the sign of $\operatorname{Im}\nu$ does not affect the proof of Proposition 2.1. However, we are not concerned with this problem, since it does not define a continuous extension of (2.16). Indeed, if $\nu \in \mathbb{C}^- \to \mu \in \mathbb{R}^+$, the solution of (2.18) has a limit (in $\mathcal{H}_{\text{loc}}$) which satisfies, instead of (2.14), the so-called incoming radiation condition

$$\lim_{R\to+\infty} \int_{\Sigma_R} |\partial_n\varphi + i\nu_0\,\varphi|^2\, d\Sigma = 0,$$

where ν_0 is, here again, the positive root of (2.12) (with $\nu = \mu$).

2.5. Reduction to a bounded domain. In this paragraph, ν denotes either a real positive number or a complex number of $\mathbb{C}^+$. Consider the vertical cylinder $\Sigma = \{x \in \Omega;\ x_1^2 + x_2^2 = r_0^2\}$, where $r_0 > 0$ is chosen large enough so that Σ does not intersect Γ. This boundary Σ splits Ω into an inner bounded part $\hat{\Omega}$ and an outer unbounded part $\check{\Omega}$ (see Fig. 2.2). The parts of S and F which are contained in the boundary $\partial\hat{\Omega}$ of $\hat{\Omega}$ (respectively, $\partial\check{\Omega}$ of $\check{\Omega}$) are denoted by $\hat{S}$ and $\hat{F}$ (respectively, $\check{S}$ and $\check{F}$).

Let us define the outer Dirichlet problem, denoted by $\check{\mathcal{P}}_\nu$ in the sequel:

$$(2.25)\qquad \begin{array}{l} \text{Find } \check{\varphi} \ \text{ in } H^1_{\text{loc}}(\check{\Omega}) \ \text{ if } \nu \in \mathbb{R}^+ \ \text{ or } \ \text{in } H^1(\check{\Omega}) \text{ if } \nu \in \mathbb{C}^+ \ \text{ such that} \\ \Delta\check{\varphi} = 0 \quad \text{in } \check{\Omega}, \\ \partial_n\check{\varphi} - \nu\,\check{\varphi} = 0 \quad \text{on } \check{S}, \\ \partial_n\check{\varphi} = 0 \quad \text{on } \check{F}, \\ \check{\varphi} = \chi \quad \text{on } \Sigma, \\ \check{\varphi} \text{ satisfies the radiation condition (2.14)} \quad \text{if } \nu \in \mathbb{R}^+, \end{array}$$

where χ is a given function defined on Σ.

PROPOSITION 2.2. *For every $\nu \in \mathbb{R}^+ \cup \mathbb{C}^+$ and for every datum $\chi \in H^{1/2}(\Sigma)$, problem $\check{\mathcal{P}}_\nu$ has a unique solution $\check{\varphi} = \check{\mathcal{R}}_\nu\,\chi$ which depends continuously on χ, i.e., if $\nu \in \mathbb{R}^+$, for every bounded domain $\mathcal{O} \subset \check{\Omega}$, there exists $K(\nu, \mathcal{O}) > 0$ such that*

$$(2.26)\qquad \left\|\check{\mathcal{R}}_\nu\,\chi\right\|_{1,\mathcal{O}} \le K(\nu,\mathcal{O})\|\chi\|_{1/2,\Sigma}\,;$$

if $\nu \in \mathbb{C}^+$, there exists $K(\nu) > 0$ such that

$$(2.27)\qquad \left\|\check{\mathcal{R}}_\nu\,\chi\right\|_{1,\check{\Omega}} \le K(\nu)\|\chi\|_{1/2,\Sigma}.$$

Proof. (i) If $\nu \in \mathbb{R}^+$, the uniqueness of the solution of $\check{\mathcal{P}}_\nu$ can be worked out by John's method [11]. For the existence proof, we may use some general methods such as the limiting absorption principle (see Remark 2.2) or the "compactness method" (which consists in writing $\check{\mathcal{P}}_\nu$ as a Fredholm equation; see, e.g., [5]). Both methods are based on an integral representation of $\check{\varphi}$ in $\check{\Omega}$. They allow us to deduce the existence of a solution from the uniqueness property. However, for the sake of consistency, we will give in the sequel (see Lemma 3.2 proved in Appendix A) a new proof of the existence and the continuity property (2.26) which is closely related to the method explained in this paper: we will see that the solution can be expressed as the sum of a converging series in $H^1_{\text{loc}}(\check{\Omega})$.

(ii) Now, we deal with the complex case. From classical trace theorems, we know that for every $\chi \in H^{1/2}(\Sigma)$, there exists $\check{\chi} \in H^1(\check{\Omega})$ (which can be chosen with compact support) such that $\check{\chi}_{|\Sigma} = \chi$ and

$$\|\check{\chi}\|_{1,\check{\Omega}} \leq C\|\chi\|_{1/2,\Sigma}, \tag{2.28}$$

where C is a positive constant. Setting $\check{\psi} = \check{\varphi}-\check{\chi}$, we are thus led to consider a problem similar to $\check{\mathcal{P}}_\nu$ with a homogeneous Dirichlet boundary condition on Σ. The end of the proof is similar to that of Proposition 2.1: writing a variational formulation of this problem in the space $\left\{\check{\psi} \in H^1(\check{\Omega});\ \check{\psi}_{|\Sigma} = 0\right\}$, the existence and the uniqueness of the solution follows from Lax–Milgram's theorem, which, moreover, gives the following inequality:

$$\|\check{\psi}\|_{1,\check{\Omega}} \leq K'(\nu)\, \|\check{\chi}\|_{1,\check{\Omega}} \cdot$$

Property (2.27) thus derives from this estimate and (2.28). □

Consider now the so-called coupling operator $\mathcal{Q}_\nu$ from $H^{1/2}(\Sigma)$ into $H^{-1/2}(\Sigma)$ given by

$$\mathcal{Q}_\nu \chi = \partial_n(\check{\mathcal{R}}_\nu \chi)_{|\Sigma}, \tag{2.29}$$

where ∂_n denotes here the exterior normal derivative to $\check{\Omega}$. Proposition 3.1 shows that, for every $\nu \in \mathbb{R}^+ \cup \mathbb{C}^+$, $\mathcal{Q}_\nu$ is a continuous operator. We then define the reduced problem, denoted by $\hat{\mathcal{P}}_\nu$:

$$\begin{array}{l}
\text{Find } \hat{X} = (\hat{\varphi}, \hat{u}) \in \hat{\mathcal{H}} \quad \text{such that} \\
\Delta\hat{\varphi} = 0 \quad \text{in } \hat{\Omega}, \\
\partial_n\hat{\varphi} - \nu\hat{\varphi} = 0 \quad \text{on } \hat{S}, \\
\partial_n\hat{\varphi} = 0 \quad \text{on } \hat{F}, \\
\nu\rho\, \hat{u}_i + \partial_{x_j}\sigma_{ij}(\hat{u}) = 0 \quad \text{in } B, \\
\sigma_{ij}(\hat{u})\, n_j = 0 \quad \text{on } \Gamma_0, \\
\nu^{1/2}\, \hat{u}.n - \partial_n\hat{\varphi} = f' \quad \text{on } \Gamma, \\
\sigma_{ij}(\hat{u})\, n_j + \nu^{1/2}\, \hat{\varphi}\, n_i = n_i f'' \quad \text{on } \Gamma, \\
\partial_n\hat{\varphi} = -\mathcal{Q}_\nu\, \hat{\varphi}_{|\Sigma} \quad \text{on } \Sigma,
\end{array} \tag{2.30}$$

where $\hat{\mathcal{H}}$ is the Hilbert space

$$\hat{\mathcal{H}} = H^1(\hat{\Omega}) \times H^1(B)^3. \tag{2.31}$$

This problem is equivalent to problem $\mathcal{P}_\nu$ in the following sense.

PROPOSITION 2.3. *Let $\nu \in \mathbb{R}^+ \cup \mathbb{C}^+$; for every $f = (f', f'') \in \mathcal{F}$, problem $\mathcal{P}_\nu$ has at least (respectively, at most) one solution if and only if the same holds for $\hat{\mathcal{P}}_\nu$.*

Moreover, if $\hat{X} = (\hat\varphi, \hat u)$ *is a solution of* $\hat{\mathcal{P}}_\nu$, *then* $X = (\varphi, \hat u)$, *where* φ *is the function given by*

$$(2.32) \qquad \varphi_{|\hat\Omega} = \hat\varphi \quad \text{and} \quad \varphi_{|\check\Omega} = \check{\mathcal{R}}_\nu\, \hat\varphi_{|\Sigma},$$

is a solution of $\mathcal{P}_\nu$. *Conversely, if* $X = (\varphi, u)$ *is a solution of* $\mathcal{P}_\nu$, *then* $\hat X = (\varphi_{|\hat\Omega}, u)$ *is a solution of* $\hat{\mathcal{P}}_\nu$.

Proof. It is enough to prove the equivalence between $\mathcal{P}_\nu$ and $\hat{\mathcal{P}}_\nu$ for the existence property : the equivalence for the uniqueness follows by contraposition from the unique continuation property. Let us first assume that $\hat{\mathcal{P}}_\nu$ admits (at least) one solution $(\hat\varphi, \hat u)$. The coupling condition $\partial_n \hat\varphi = -\mathcal{Q}_\nu\, \hat\varphi_{|\Sigma}$ implies the analytical matching on Σ between $\hat\varphi$ and $\check{\mathcal{R}}_\nu\, \hat\varphi_{|\Sigma}$; according to (2.25) and (2.30) of $\check{\mathcal{P}}_\nu$ and $\hat{\mathcal{P}}_\nu$, the pair $(\varphi, \hat u)$ (where φ is given by (2.32)) actually satisfies the equations of $\mathcal{P}_\nu$. Conversely, if (φ, u) is a solution of $\mathcal{P}_\nu$, the restriction to $\check\Omega$ of φ is nothing but $\check{\mathcal{R}}_\nu\, \varphi_{|\Sigma}$, and thus $\hat\varphi = \varphi_{|\hat\Omega}$ clearly satisfies the coupling condition on Σ. □

Remark 2.4. Note that if $\nu \in \mathbb{C}^+$, Propositions 2.1 and 2.3 show that problem $\hat{\mathcal{P}}_\nu$ is well posed.

3. Diagonalization of the coupling operator. The purpose of this section is to express the coupling operator $\mathcal{Q}_\nu$ (given by (2.29)) by means of an explicit series expansion. We will see that for a suitable choice of a basis of $H^{1/2}(\Sigma)$ (which depends on ν), $\mathcal{Q}_\nu$ becomes diagonal. This result is based on the method of separation of variables which is used here for the outer problem in cylindrical coordinates (r, θ, x_3) (with $r = (x_1^2 + x_2^2)^{1/2}$). In §3.1, we first deal with the case of real positive ν for which the main results follow from spectral theory. In §3.2, we generalize these results to complex $\nu \in \mathbb{C}^+$, using perturbation techniques. We finally show in §3.3 that $\mathcal{Q}_\nu$ actually depends analytically on ν, and has an analytic continuation in the lower complex half plane. This latter property is the basic tool for the analytic extension of the reduced problem (§4).

3.1. The case of real positive ν. The method we use here is quite similar to the one introduced by Lenoir and Tounsi [17] in the two-dimensional case. We first exhibit the solutions of $\check{\mathcal{P}}_\nu$ with separated variables. We then prove that every solution of $\check{\mathcal{P}}_\nu$ can be expanded as a series in terms of solutions with separated variables.

In the following lemma, K_n denotes the modified Bessel function, which is real-valued and exponentially decreasing at infinity for real positive arguments (see, e.g., [1]).

LEMMA 3.1. *Let* $\nu > 0$; *the solutions with separated variables of the outer problem* $\check{\mathcal{P}}_\nu$ *are given (up to a multiplicative complex constant) by*

$$(3.1) \qquad \check\varphi_\nu^{(n,m)}(r, \theta, x_3) = \eta_\nu^{(n,m)}(r)\, \chi_\nu^{(n,m)}(\theta, x_3) \quad \text{for } n \in \mathbb{Z} \quad \text{and} \quad m \geq 0,$$

where the functions $\eta_\nu^{(n,m)}$ *and* $\chi_\nu^{(n,m)}$ *are defined as follows:*

$$(3.2) \quad \eta_\nu^{(n,m)}(r) = \frac{K_n(\zeta_\nu^{(m)} r)}{K_n(\zeta_\nu^{(m)} r_0)}, \qquad \chi_\nu^{(n,m)}(\theta, x_3) = \frac{e^{in\theta}}{\sqrt{2\pi r_0}}\, \tau_\nu^{(m)}(x_3) \quad \text{and}$$

$$(3.3) \quad \tau_\nu^{(m)}(x_3) = \alpha_\nu^{(m)} \cos(\zeta_\nu^{(m)}(x_3 + 1)) \quad \text{with } \alpha_\nu^{(m)} = \left(\frac{\sin 2\zeta_\nu^{(m)}}{4\zeta_\nu^{(m)}} + \frac{1}{2} \right)^{-1/2},$$

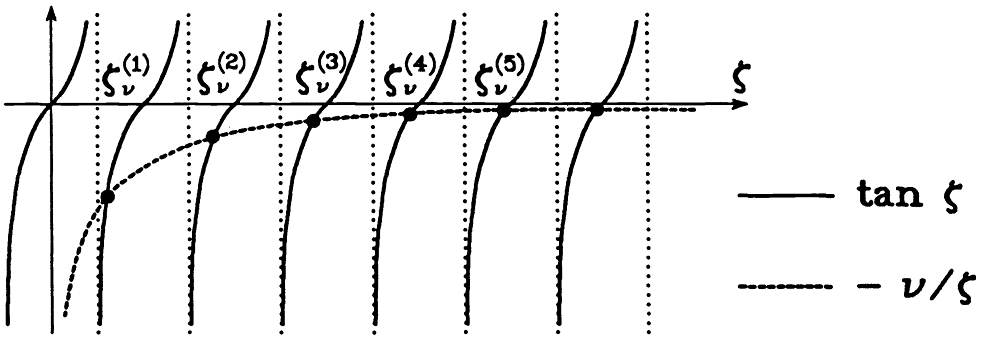


FIG. 3.1. *Real solutions of the dispersion equation.*

and the $\zeta_\nu^{(m)}$, $m \geq 0$, *are the roots of the so-called dispersion equation*

$$-\zeta \tan \zeta = \nu, \tag{3.4}$$

chosen such that $\zeta_\nu^{(0)}$ *is imaginary with negative imaginary part, and* $\zeta_\nu^{(m)}$, *for* $m \geq 1$, *are real, positive, and arranged according to increasing values.*

Remark 3.1. The positive roots $\zeta_\nu^{(m)}$, $m \geq 1$, of (3.4) are represented in Fig. 3.1 ; it may be easily seen that

$$\zeta_\nu^{(m)} \in \,](m - 1/2)\pi, m\pi[\quad \text{for } m \geq 1 \quad \text{and} \quad \zeta_\nu^{(m)} \sim m\pi \quad \text{when } m \to +\infty. \tag{3.5}$$

Notice that (3.4) has only one imaginary root with negative imaginary part. Indeed, setting $\zeta_\nu^{(0)} = -i\nu_0$, we see that ν_0 is the only positive root of (2.12) which appears in the radiation condition (2.14).

Remark 3.2. In this definition of $\check{\varphi}_\nu^{(n,m)}$, the coefficients are chosen so that $\check{\varphi}_\nu^{(n,m)}$ is the (only) solution of $\check{\mathcal{P}}_\nu$ for the Dirichlet datum $\chi_\nu^{(n,m)}$ (which can be considered as a function defined on Σ), and this latter function is a unit vector of $L^2(\Sigma)$. Moreover, setting

$$\Sigma = \mathcal{C} \times \mathcal{I}, \tag{3.6}$$

where $\mathcal{C}$ is a circle of radius r_0 and $\mathcal{I}$ is the vertical segment $]-1, 0[$, we see that functions $\tau_\nu^{(m)}$ and $e^{in\theta}/\sqrt{2\pi r_0}$ are unitary, respectively, in $L^2(\mathcal{I})$ and $L^2(\mathcal{C})$.

Remark 3.3. If $H_n^{(1)}$ denotes the Hankel function of order n of the first kind, we have (see [1])

$$K_n(z) = \frac{i\pi}{2}\, e^{in\pi/2}\, H_n^{(1)}(iz) \quad \text{for } -\pi < \arg z \leq \frac{\pi}{2}. \tag{3.7}$$

Hence, from Remark 3.1, we see that the definitions of $\eta_\nu^{(n,0)}$ and $\chi_\nu^{(n,0)}$ amount to

$$\eta_\nu^{(n,0)}(r) = \frac{H_n^{(1)}(\nu_0 r)}{H_n^{(1)}(\nu_0 r_0)}, \qquad \chi_\nu^{(n,0)}(\theta, x_3) = \frac{e^{in\theta}}{\sqrt{2\pi r_0}}\,\tau_\nu^{(0)}(x_3) \quad \text{and} \tag{3.8}$$

$$\tau_\nu^{(0)}(x_3) = \alpha_\nu^{(0)} \cosh(\nu_0(x_3 + 1)) \quad \text{with } \alpha_\nu^{(0)} = \left(\frac{\sinh 2\nu_0}{4\nu_0} + \frac{1}{2}\right)^{-1/2}. \tag{3.9}$$

These expressions clearly show that the solutions $\check{\varphi}_\nu^{(n,m)}$ of $\check{\mathcal{P}}_\nu$ are divided into two classes. If $m > 0$, $\check{\varphi}_\nu^{(n,m)}$ decreases exponentially when $r \to +\infty$, these solutions are called "evanescent." But if $m = 0$, the functions $\check{\varphi}_\nu^{(n,0)}$ decrease as $r^{-1/2}$ (see (3.14) below), they are called "radiative."

Remark 3.4. Note that, from the relation

$$K_{-n}(z) = K_n(z), \quad n \in \mathbb{Z}, \quad \text{and} \quad |\arg z| < \pi, \tag{3.10}$$

we have the symmetry properties

$$\eta_\nu^{(-n,m)} = \eta_\nu^{(n,m)} \quad \text{and} \quad \chi_\nu^{(-n,m)} = \overline{\chi_\nu^{(n,m)}}. \tag{3.11}$$

Proof of Lemma 3.1. Suppose that a solution $\check{\varphi}$ of $\check{\mathcal{P}}_\nu$ can be written in the form

$$\check{\varphi}(r,\theta,x_3) = \eta(r)\,\tau(x_3)\,\frac{e^{in\theta}}{\sqrt{2\pi r_0}},$$

where $n \in \mathbb{Z}$. According to (2.25) of $\check{\mathcal{P}}_\nu$, functions $\eta(r)$ and $\tau(x_3)$ are solutions of the following problems for some $\lambda \in \mathbb{R}$:

$$\begin{array}{l} r d_r(r d_r \eta) - (n^2 + \lambda r^2)\eta = 0 \quad \text{on }]r_0, +\infty[, \\ \lim_{R\to+\infty} R\,|d_r\eta(R) - i\nu_0\,\eta(R)|^2 = 0, \end{array} \tag{3.12}$$

and

$$\begin{array}{l} d_{x_3}^2\tau + \lambda\tau = 0 \quad \text{on } \mathcal{I}, \\ d_{x_3}\tau(0) - \nu\,\tau(0) = 0, \\ d_{x_3}\tau(-1) = 0 \end{array} \tag{3.13}$$

(where $\mathcal{I}$ denotes the segment $]-1,0[$). Solving the eigenvalue problem (3.13) does not raise any difficulty. If $\lambda > 0$, we obtain the sequence of solutions $(\lambda_\nu^{(m)}, \tau_\nu^{(m)})$ with $\lambda_\nu^{(m)} = {\zeta_\nu^{(m)}}^2$, $m \geq 1$. On the other hand, if $\lambda < 0$, problem (3.13) has a unique solution $(\lambda_\nu^{(0)}, \tau_\nu^{(0)})$, where $\lambda_\nu^{(0)} = -\nu_0^2$.

We can then solve problem (3.12). If $\lambda = {\zeta_\nu^{(m)}}^2$ with $m \geq 1$, the solution can be expressed by means of the modified Bessel function K_n and I_n of order n (see [1]):

$$\eta(r) = \gamma\, K_n(\zeta_\nu^{(m)} r) + \gamma'\, I_n(\zeta_\nu^{(m)} r), \qquad (\gamma,\gamma') \in \mathbb{C}^2,$$

where γ' must be zero, since I_n increases exponentially at infinity. If $\lambda = -\nu_0^2$, the solution is a combination of the Hankel functions $H_n^{(1)}(\nu_0 r)$ and $H_n^{(2)}(\nu_0 r)$; from the asymptotic behaviours of $H_n^{(j)}(z)$ $(j = 1,2)$ and their derivatives for fixed $n \in \mathbb{Z}$ and large $z \in \mathbb{R}^+$ (see [1])

$$H_n^{(j)}(z) = \sqrt{\frac{2}{\pi z}}\, e^{(-1)^{j-1} i(z - n\pi/2 - \pi/4)} + O(z^{-3/2}), \tag{3.14}$$

$$H_n^{(j)\prime}(z) = (-1)^{j-1} i H_n^{(j)}(z) + O(z^{-3/2}), \tag{3.15}$$

we deduce that only $H_n^{(1)}(\nu_0 r)$ satisfies the radiation condition in (3.12). □

PROPOSITION 3.1. *Let $\nu > 0$; the set $\mathcal{X}_\nu = \{\chi_\nu^{(n,m)};\ n \in \mathbb{Z},\ m \geq 0\}$ is an orthonormal basis of $L^2(\Sigma)$ and an orthogonal basis of $H^s(\Sigma)$ for every $s \in\,]0,1]$. Furthermore, the expression*

$$[\chi]_{\nu,s} = \left(\sum_{n\in\mathbb{Z}} \sum_{m\geq 0} (1 + n^2 + m^2)^s \left| (\chi, \chi_\nu^{(n,m)})_\Sigma \right|^2 \right)^{1/2} \tag{3.16}$$

is a norm on $H^s(\Sigma)$ which is equivalent to the usual norm.

Proof. (i) Let us prove that the functions $\chi_\nu^{(m,n)}$ are the eigenvectors of a selfadjoint operator with compact resolvent. First, notice that each function $\chi_\nu^{(n,m)}$ is a solution of the following eigenvalue problem for $\lambda = r_0^{-2}n^2 + (\zeta_\nu^{(m)})^2$:

$$(3.17) \qquad \begin{aligned} &\Delta_\sigma \chi + \lambda \chi = 0 \quad \text{in } \Sigma, \\ &\partial_{x_3}\chi - \nu\chi = 0 \quad \text{on } \mathcal{C}_0, \\ &\partial_{x_3}\chi = 0 \quad \text{on } \mathcal{C}_{-1}, \end{aligned}$$

where Δ_σ denotes the Laplace–Beltrami operator on Σ (i.e., $r_0^{-2}\partial_\theta^2 + \partial_{x_3}^2$), $\mathcal{C}_0$ and $\mathcal{C}_{-1}$ are, respectively, the upper and lower boundaries of Σ (i.e., two circles of radius r_0). A variational formulation of this problem writes as follows:

$$\begin{aligned} &\text{Find } \chi \in H^1(\Sigma) \text{ such that} \\ &t_\nu(\chi,\chi') = \lambda\,(\chi,\chi')_\Sigma \quad \forall \chi' \in H^1(\Sigma), \end{aligned}$$

where $t_\nu(\cdot,\cdot)$ is the following symmetric sesquilinear form on $H^1(\Sigma)$ (∇_σ denotes the tangential gradient on Σ):

$$t_\nu(\chi,\chi') = \int_\Sigma \nabla_\sigma\chi.\overline{\nabla_\sigma\chi'}\,d\Sigma - \nu\int_{\mathcal{C}_0}\chi\overline{\chi'}\,d\mathcal{C}.$$

From Lions's lemmas [20] we know that, for any $\varepsilon > 0$, there exists $K_\varepsilon > 0$ such that

$$C\,\|\chi\|_{\mathcal{C}_0} \le \|\chi\|_{2/3,\Sigma} \le \varepsilon\|\chi\|_{1,\Sigma} + K_\varepsilon\|\chi\|_\Sigma \quad \forall\chi \in H^1(\Sigma),$$

from which we infer that, for sufficiently large $d > 0$, the form

$$t_{d,\nu}(\chi,\chi') = t_\nu(\chi,\chi') + d\,(\chi,\chi')_\Sigma$$

is coercive on $H^1(\Sigma)$. As a consequence, the representation theorem of sesquilinear forms (see, e.g., Reed and Simon [27]) shows that the operator $T_{d,\nu}$ defined on $L^2(\Sigma)$ by

$$\begin{aligned} &D(T_{d,\nu}) = \{\chi \in H^1(\Sigma);\ \exists K > 0,\ \forall\chi' \in H^1(\Sigma),\ t_{d,\nu}(\chi,\chi') \le K\|\chi'\|_\Sigma\}, \\ &(T_{d,\nu}\,\chi,\chi')_\Sigma = t_{d,\nu}(\chi,\chi') \quad \forall\chi \in D(T_{d,\nu}) \quad \forall\chi' \in H^1(\Sigma), \end{aligned}$$

is an unbounded selfadjoint positive operator with compact resolvent. Thus $T_{d,\nu}$ has a countable infinity of positive eigenvalues and the associated eigenvectors can be chosen so as to form an orthonormal basis of $L^2(\Sigma)$.

(ii) To see that the functions $\chi_\nu^{(n,m)}$ are the only eigenvectors of $T_{d,\nu}$, first notice that each of the two families $\{e^{in\theta}/\sqrt{2\pi r_0};\ n \in \mathbb{Z}\}$ and $\{\tau_\nu^{(m)};\ m \ge 0\}$ is, respectively, an orthonormal basis of $L^2(\mathcal{C})$ and $L^2(\mathcal{I})$. The former result is classical. To prove the latter, it suffices to apply the method described in (i) to the eigenvalue problem (3.13) (instead of (3.17)): since we have determined all the solutions of this one-dimensional problem, the completeness of the $\tau_\nu^{(m)}$ follows. Consequently, the products $\tau_\nu^{(m)}(x_3)\,e^{in\theta}/\sqrt{2\pi r_0}$, for $n \in \mathbb{Z}$ and $m \ge 0$, form an orthonormal basis of $L^2(\Sigma)$ (see, e.g., Kato [13, Ex. V.1.10]).

(iii) The set $\{\chi_\nu^{(n,m)};\ n \in \mathbb{Z},\ m \ge 0\}$ is also an orthogonal basis of the domain $D(T_{d,\nu}^r)$ of any power $T_{d,\nu}^r$ (with $r > 0$) of operator $T_{d,\nu}$:

$$\begin{aligned} &D(T_{d,\nu}^r) = \left\{\chi \in L^2(\Sigma);\ \textstyle\sum_{n\in\mathbb{Z}}\sum_{m\ge0}(\lambda_\nu^{(n,m)})^{2r}\left|\left(\chi,\chi_\nu^{(n,m)}\right)_\Sigma\right|^2 < \infty\right\}, \\ &T_{d,\nu}^r\,\chi = \textstyle\sum_{n\in\mathbb{Z}}\sum_{m\ge0}(\lambda_\nu^{(n,m)})^r\left(\chi,\chi_\nu^{(n,m)}\right)_\Sigma\chi_\nu^{(n,m)} \quad \forall\chi \in D(T_{d,\nu}^r), \end{aligned}$$

where $\lambda_\nu^{(n,m)} = d + r_0^{-2}n^2 + \zeta_\nu^{(m)^2}$. By the interpolation theory between Sobolev spaces (see Lions and Magenes [21]), we see that $D(T^r_{d,\nu})$, for $r \in [0, \frac{1}{2}]$, is nothing but the interpolation space $[H^1(\Sigma), L^2(\Sigma)]_{1-2r}$, and thus

$$\mathrm{D}(T^r_{d,\nu}) = H^{2r}(\Sigma) \quad \text{for } 0 \le r \le \tfrac{1}{2}. \tag{3.18}$$

Consequently, $\mathcal{X}_\nu$ is an orthogonal basis of $H^s(\Sigma)$ for $s \in [0,1]$ and the expression

$$\left(\sum_{n\in\mathbb{Z}} \sum_{m\geq 0} (\lambda_\nu^{(n,m)})^s \left| \left(\chi, \chi_\nu^{(n,m)} \right)_\Sigma \right|^2 \right)^{1/2}$$

is equivalent to the usual norm of $H^s(\Sigma)$. Moreover, we have from (3.5),

$$C_1 \left(d + r_0^{-2}n^2 + \zeta_\nu^{(m)^2} \right) \le (1 + n^2 + m^2) \le C_2 \left(d + r_0^{-2}n^2 + \zeta_\nu^{(m)^2} \right)$$

for some positive constants C_1 and C_2. The statement of the proposition follows. □

Proposition 3.1 shows that every function $\chi \in H^s(\Sigma)$ (with $0 \le s \le 1$) expands as

$$\chi = \sum_{n\in\mathbb{Z}} \sum_{m\geq 0} (\chi, \chi_\nu^{(n,m)})_\Sigma \, \chi_\nu^{(n,m)}, \tag{3.19}$$

where the series converges in $H^s(\Sigma)$. The following result completes the proof of Proposition 2.2 (existence of a solution of $\check{\mathcal{P}}_\nu$ for real positive ν).

LEMMA 3.2. *Let $\nu > 0$; for every $\chi \in H^{1/2}(\Sigma)$, the outer problem $\check{\mathcal{P}}_\nu$ has a unique solution $\check{\varphi} = \check{\mathcal{R}}_\nu \chi$ which expands as*

$$\check{\varphi} = \sum_{n\in\mathbb{Z}} \sum_{m\geq 0} (\chi, \chi_\nu^{(n,m)})_\Sigma \, \check{\varphi}_\nu^{(n,m)}, \tag{3.20}$$

where the series converges in $H^1_{\mathrm{loc}}(\check{\Omega})$ and depends continously on χ.

Proof. Recall that the (only) solution of $\check{\mathcal{P}}_\nu$ for the Dirichlet datum $\chi_\nu^{(n,m)}$ is $\check{\varphi}_\nu^{(n,m)}$ (defined in Lemma 3.1). Consequently, if the series (3.20) converges in $H^1_{\mathrm{loc}}(\check{\Omega})$ and satisfies the radiation condition, it will clearly define the only solution of $\check{\mathcal{P}}_\nu$ for the Dirichlet datum χ. The proof of this statement (which is rather technical) and of the continuity with respect to χ is given in Appendix A. □

Finally, by (2.26) and the continuity of the normal derivative of $\check{\varphi}$ on Σ, we deduce the diagonal form of the coupling operator.

COROLLARY 3.1. *Let $\nu > 0$; the coupling operator $\mathcal{Q}_\nu$ given by (2.29) expands as*

$$\mathcal{Q}_\nu \chi = \sum_{n\in\mathbb{Z}} \sum_{m\geq 0} q_\nu^{(n,m)} \, (\chi, \chi_\nu^{(n,m)})_\Sigma \, \chi_\nu^{(n,m)} \quad \textit{with } q_\nu^{(n,m)} = -\zeta_\nu^{(m)} \frac{K_n'(\zeta_\nu^{(m)} r_0)}{K_n(\zeta_\nu^{(m)} r_0)}, \tag{3.21}$$

where the series converges in $H^{-1/2}(\Sigma)$ for every $\chi \in H^{1/2}(\Sigma)$.

3.2. The case of complex ν. The extension of the results of §3.1 to complex values of ν raises two difficulties. On one hand, we have to solve the dispersion equation (3.4) in the complex plane. On the other hand, the eigenvalue problem (3.17) is no

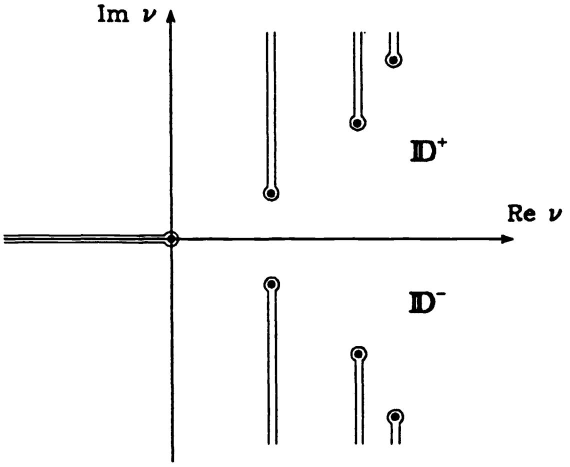


FIG. 3.2. *Analyticity domain of the* $\zeta_\nu^{(m)}$.

longer selfadjoint and therefore the spectral theory results we have used for real ν do not apply any more.

For the sake of clarity, the study of the dispersion equation is postponed to Appendix B. We show that each root $\zeta_\nu^{(m)}$, for $m \geq 0$, actually extends to an analytic function of ν (which will still be denoted by $\zeta_\nu^{(m)}$) in every simply connected domain of $\mathbb{C} \setminus \mathbb{E}$, where $\mathbb{E}$ is the discrete set of $\mathbb{C}$ (with no accumulation point) which consists of the values of ν for which the dispersion equation admits a double root (see §B.1). In the sequel, we will consider the simply connected domain $\mathbb{D}$ of $\mathbb{C} \setminus \mathbb{E}$ which is obtained by cutting the complex plane from each point of $\mathbb{E}$ as shown in Fig. 3.2. This is of course an arbitrary choice, which is justified here by the fact that we are primarily concerned with the analytic continuation of the problem in the vicinity of the positive real axis. We will denote $\mathbb{D}^\pm = \mathbb{D} \cap \mathbb{C}^\pm$.

Since each solution $\zeta_\nu^{(m)}$, for $m \geq 0$, depends analytically on ν in $\mathbb{D}$, each function $\tau_\nu^{(m)}$ given by (3.3) (and thus $\chi_\nu^{(n,m)}$ defined in (3.2)) can be extended to an analytic family on $\mathbb{D}$. Note that the quantity $\sin(2\zeta_\nu^{(m)}) + 2\zeta_\nu^{(m)}$ does not vanish if $\nu \in \mathbb{C} \setminus \mathbb{E}$ (see Proposition B.1), which shows that the normalization coefficient $\alpha_\nu^{(m)}$ is always defined in this domain.

Similarly, each function $\check{\varphi}_\nu^{(n,m)}$can be extended analytically in $\mathbb{D}$ except maybe for the values of ν for which $K_n(\zeta_\nu^{(m)} r_0) = 0$. If $\nu \in \mathbb{D}^+$, this quantity cannot vanish: this results from property (B.34) and the fact that the modified Bessel functions K_n have no zeros in the region $|\arg z| \leq \pi/2$. Thus, $\check{\varphi}_\nu^{(n,m)}$ is defined everywhere in $\mathbb{D}^+$: it is obviously the only solution of $\check{\mathcal{P}}_\nu$ for the Dirichlet datum $\chi_\nu^{(n,m)}$. From (B.34), we see that it is always exponentially decreasing when $r \to +\infty$.

The purpose of this paragraph is to show that Proposition 3.1 and Corollary 3.1 extend as follows.

THEOREM 3.1. *For every $\nu \in \mathbb{D}$ and every $s \in [0,1]$, the set $\mathcal{X}_\nu = \{\chi_\nu^{(n,m)};\ n \in \mathbb{Z}, m \geq 0\}$ is a basis of $H^s(\Sigma)$, and the expression $[\![\chi]\!]_{\nu,s}$ given by* (3.16) *is still a norm on $H^s(\Sigma)$ equivalent to the usual norm. Moreover, the two families $\mathcal{X}_\nu$ and $\mathcal{X}_{\bar\nu}$ are adjoint to each other in $L^2(\Sigma)$, i.e.,*

$$(\chi_\nu^{(n,m)}, \chi_{\bar\nu}^{(n',m')})_\Sigma = \delta_{nn'}\delta_{mm'}. \tag{3.22}$$

In the particular case $\nu \in \mathbb{D}^+$, this statement allows us to extend the diagonal form (3.21) of the coupling operator (which was up to now defined for $\nu \in \mathbb{R}^+ \cup \mathbb{C}^+$) as follows.

COROLLARY 3.2. *Let $\nu \in \mathbb{D}^+$; the coupling operator $\mathcal{Q}_\nu$ given by (2.29) expands as*

$$\mathcal{Q}_\nu \chi = \sum_{n\in\mathbb{Z}} \sum_{m\geq 0} q_\nu^{(n,m)} \, (\chi, \chi_{\bar\nu}^{(n,m)})_\Sigma \, \chi_\nu^{(n,m)} \quad \textit{with } q_\nu^{(n,m)} = -\zeta_\nu^{(m)} \frac{K_n'(\zeta_\nu^{(m)} r_0)}{K_n(\zeta_\nu^{(m)} r_0)}, \tag{3.23}$$

where the series converges in $H^{-1/2}(\Sigma)$ for every $\chi \in H^{1/2}(\Sigma)$.

Note that this expression of $\mathcal{Q}_\nu$ agrees with (3.21) if $\nu \in \mathbb{R}^+$, since $\nu = \bar\nu$ in this case.

Proof of Corollary 3.2. Theorem 3.1 shows that every function $\chi \in H^{1/2}(\Sigma)$ expands as

$$\chi = \sum_{n\in\mathbb{Z}} \sum_{m\geq 0} (\chi, \chi_{\bar\nu}^{(n,m)})_\Sigma \, \chi_\nu^{(n,m)}, \tag{3.24}$$

where the series converges in $H^{1/2}(\Sigma)$. Since $\check\varphi_\nu^{(n,m)}$ is the solution of $\check{\mathcal{P}}_\nu$ for the datum $\chi_\nu^{(n,m)}$, we deduce from the continuity of $\check{\mathcal{R}}_\nu$ (Lemma 3.2) that the general solution of $\check{\mathcal{P}}_\nu$ is given by

$$\check\varphi = \check{\mathcal{R}}_\nu \chi = \sum_{n\in\mathbb{Z}} \sum_{m\geq 0} (\chi, \chi_{\bar\nu}^{(n,m)})_\Sigma \, \check\varphi_\nu^{(n,m)}, \tag{3.25}$$

where the series converges in $H^1(\check\Omega)$. The expression of $\mathcal{Q}_\nu$ follows. □

The remainder of the present paragraph is devoted to the proof of Theorem 3.1: it is based on perturbation techniques which consist in comparing the family $\mathcal{X}_\nu$ with the orthogonal basis $\mathcal{X}_\mu$ for some given $\mu > 0$. We begin by proving the completeness of the family $\mathcal{X}_\nu$ in $L^2(\Sigma)$ (Lemma 3.4) which follows from the completeness of the $\tau_\nu^{(m)}$ in $L^2(\mathcal{I})$ (Lemma 3.3) by separation of variables. Lemma 3.5 provides a perturbation result which is used in the proofs of Lemma 3.3 (in the simple case $s = 0$) and Theorem 3.1 (in the general case).

LEMMA 3.3. *For every $\nu \in \mathbb{D}$, the set $\mathcal{T}_\nu = \{\tau_\nu^{(m)};\ m \geq 0\}$ is a basis of $L^2(\mathcal{I})$ and the two families $\mathcal{T}_\nu$ and $\mathcal{T}_{\bar\nu}$ are adjoint to each other, i.e.,*

$$(\tau_\nu^{(m)}, \tau_{\bar\nu}^{(m')})_\mathcal{I} = \delta_{mm'}. \tag{3.26}$$

Remark 3.5. For real positive ν, we defined $\zeta_\nu^{(0)}$ as an imaginary number and $\zeta_\nu^{(m)}$, for $m \geq 1$, as real positive numbers. Consequently, the analyticity property of $\zeta_\nu^{(m)}$ shows that if $\nu \in \mathbb{D}$, we have

$$\zeta_{\bar\nu}^{(0)} = -\overline{\zeta_\nu^{(0)}} \quad \text{and} \quad \zeta_{\bar\nu}^{(m)} = \overline{\zeta_\nu^{(m)}} \quad \text{for } m \geq 1, \tag{3.27}$$

from which we deduce that

$$\tau_{\bar\nu}^{(m)} = \overline{\tau_\nu^{(m)}} \quad \text{for } m \geq 0. \tag{3.28}$$

Proof of Lemma 3.3. (i) Relation (3.26) can be obtained by a direct calculation of the scalar product, noticing that

$$\int_{-1}^{0} \cos^2 \zeta(x_3+1)\,dx_3 = \frac{\sin 2\zeta}{4\zeta} + \frac{1}{2}, \tag{3.29}$$

$$\int_{-1}^{0} \cos\zeta(x_3+1)\cos\zeta'(x_3+1)\,dx_3 = \frac{\cos\zeta\cos\zeta'\ (\zeta\tan\zeta - \zeta'\tan\zeta')}{(\zeta+\zeta')(\zeta-\zeta')}. \tag{3.30}$$

We can also refer to the analyticity of $\tau_\nu^{(m)}$ with respect to ν. As a matter of fact, by the unique continuation property, (3.26) is nothing but the extension to $\mathbb{D}$ of the orthonormality relation of the $\tau_\nu^{(m)}$ for real ν.

(ii) Let μ be a given real positive number, and let $\mathcal{T}_\mu = \{\tau_\mu^{(m)};\ m \geq 0\}$. We have seen in the proof of Proposition 3.1 (point (ii)) that family $\mathcal{T}_\mu$ is an orthonormal basis of $L^2(\mathcal{I})$. To prove that $\mathcal{T}_\nu$ is a basis of $L^2(\mathcal{I})$, we use a perturbation result of orthonormal families (Kato [13, Thm. V-2.20]), by considering $\mathcal{T}_\nu$ as a perturbation of the orthonormal basis $\mathcal{T}_\mu$. We have to verify the two following conditions:

(α) there exists a positive constant C such that

$$\sum_{m\geq 0} \left\| \tau_\nu^{(m)} - \tau_\mu^{(m)} \right\|_{\mathcal{I}}^2 \leq C;$$

(β) for any sequence (a_m; $m \geq 0$) of complex numbers,

$$\tau = \sum_{m\geq 0} a_m \tau_\nu^{(m)} = 0 \ \text{ implies } \ a_m = 0 \quad \forall m \geq 0.$$

Condition (α) follows from Lemma 3.5 below in the case $s = 0$ (the quantity involved in (α) is nothing but $A_{\nu,0}$). Condition (β) is a straightforward consequence of (3.26), since $a_m = (\tau, \tau_{\bar\nu}^{(m)})$. □

LEMMA 3.4. *For every $\nu \in \mathbb{D}$, the set $\mathcal{X}_\nu$ is a basis of $L^2(\Sigma)$, and the two families $\mathcal{X}_\nu$ and $\mathcal{X}_{\bar\nu}$ are adjoint to each other, i.e.,* (3.22).

Remark 3.6. This lemma shows in particular that the scalar product in $L^2(\Sigma)$ can be expressed as follows:

$$(\chi, \chi')_\Sigma = \sum_{n\in\mathbb{Z}} \sum_{m\geq 0} (\chi, \chi_{\bar\nu}^{(n,m)})_\Sigma\, (\chi_\nu^{(n,m)}, \chi')_\Sigma. \tag{3.31}$$

Proof of Lemma 3.4. Relation (3.22) results from (3.26) by Fubini's theorem. Proceeding as in Kato [13, Ex. V-1.10]), we easily prove that $(\chi, \chi_\nu^{(n,m)})_\Sigma = 0$, for all $n \in \mathbb{Z}$ and $m \geq 0$, implies $\chi = 0$: this derives from Lemma 3.3 and the completeness of the family $\{e^{in\theta}/\sqrt{2\pi r_0};\ n \in \mathbb{Z}\}$ in $L^2(\mathcal{C})$. The completeness of $\mathcal{X}_\nu$ follows. Thus, every $\chi \in L^2(\Sigma)$ expands as

$$\chi = \sum_{n\in\mathbb{Z}} \sum_{m\geq 0} a_{n,m}\, \chi_\nu^{(n,m)},$$

where $a_{n,m} = (\chi, \chi_{\bar\nu}^{(n,m)})_\Sigma$ by virtue of (3.22). □

LEMMA 3.5. *Let μ be a given real positive number. Then, for every $s \in [-1,+1]$ and every $\nu \in \mathbb{D}$, the following quantity is finite:*

$$A_{\nu,s} = \sup_{n\in\mathbb{Z}} \sum_{m\geq 0} \sum_{m'\geq 0} \left(\frac{1+n^2+m'^2}{1+n^2+m^2}\right)^s \left|\left(\tau_\nu^{(m)} - \tau_\mu^{(m)}, \tau_\mu^{(m')}\right)_{\mathcal{I}}\right|^2 . \tag{3.32}$$

Proof. This result is based on the asymptotic behaviour of the roots $\zeta_\nu^{(m)}$ of the dispersion equation when $m \to +\infty$. We prove in Appendix B (see Proposition B.6) that

$$\zeta_\nu^{(m)} = m\pi - m^{-1}\nu/\pi + O(m^{-3}) \quad \text{when } m \to +\infty. \tag{3.33}$$

Using the definition (3.3) of $\tau_\nu^{(m)}$ and formula (3.30), we deduce that

$$\left(\tau_\nu^{(m)} - \tau_\mu^{(m)}, \tau_\mu^{(m)}\right)_{\mathcal{I}} = O(m^{-2}), \tag{3.34}$$

$$\left(\tau_\nu^{(m)} - \tau_\mu^{(m)}, \tau_\mu^{(m')}\right)_{\mathcal{I}} = \left(\tau_\nu^{(m)}, \tau_\mu^{(m')}\right)_{\mathcal{I}} = O\left(|m^2 - m'^2|^{-1}\right) \quad \text{if } m \neq m', \tag{3.35}$$

when $m^2 + m'^2 \to +\infty$. In order to prove that $A_{\nu,s}$ is bounded, we split the series into five parts $A_{\nu,s}^{(j)}$, for $j = -2$ to $+2$, which correspond, respectively, to $m' \leq am$, $am < m' < m$, $m' = m$, $m < m' < bm$, and $bm \leq m'$, where a and b are given real positive constants such that $a < 1 < b$. We will assume that $s \geq 0$: the case of negative s can be dealt with similarly. For $A_{\nu,s}^{(-2)}$ to $A_{\nu,s}^{(1)}$, we first notice that

$$\frac{1+n^2+m'^2}{1+n^2+m^2} \leq C,$$

which shows that the convergence does not depend on n and s in these cases. Using (3.34) and (3.35), we deduce

$$A_{\nu,s}^{(-2)} \leq C \sum_{m>0} \sum_{0\leq m'\leq am} (m^2 - m'^2)^{-2} \leq C \sum_{m>0} m^{-3},$$

$$A_{\nu,s}^{(-1)} \leq C \sum_{m>0} \sum_{am<m'<m} (m^2 - m'^2)^{-2} \leq C \sum_{m>0} \sum_{0<p<(1-a)m} p^{-2} m^{-2},$$

$$A_{\nu,s}^{(0)} \leq C \sum_{m>0} m^{-4},$$

and, for $A_{\nu,s}^{(1)}$, the same kind of estimate as for $A_{\nu,s}^{(-1)}$: the convergence of these series follows. It remains to deal with $A_{\nu,s}^{(2)}$. We use in this case the following inequalities (for $m' \geq bm$):

$$\frac{1+n^2+m'^2}{1+n^2+m^2} \leq \frac{m'^2}{m^2},$$

$$m'^2 - m^2 \geq C\, m^{1-\alpha} m'^{1+\alpha},$$

where α can be chosen in the range $[0,1]$. Thus, we deduce

$$A_{\nu,s}^{(2)} \leq C \sum_{m>0} \sum_{m'\geq bm} m^{-2(1-\alpha+s)} m'^{-2(1+\alpha-s)},$$

which converges if $|s-\alpha| < \frac{1}{2}$: for every $s \in [0,1]$, we can choose α such that this condition is satisfied. □

We can now proceed to the proof of Theorem 3.1. As for Lemma 3.3, we consider $\mathcal{X}_\nu$ as a perturbation of the orthogonal family $\mathcal{X}_\mu$ for a given $\mu \in \mathbb{R}^+$. However, the perturbation result we have used in the proof of that lemma does not apply here: condition (α) is not satisfied, since for every $m \geq 0$, the quantity $\sum_{n\in\mathbb{Z}}[\chi_\nu^{(n,m)} - \chi_\mu^{(n,m)}]^2_{\mu,s}$ is infinite. In other words, the perturbation only concerns function $\tau_\nu^{(m)}$ (which does not depend on n) in the expression of $\chi_\nu^{(n,m)}$. Namely, we have

$$\left(\chi_\nu^{(n,m)}, \chi_\mu^{(n',m')}\right)_\Sigma = \delta_{nn'} \left(\tau_\nu^{(m)}, \tau_\mu^{(m')}\right)_{\mathcal{I}}. \tag{3.36}$$

Proof of Theorem 3.1. Consider the linear operators T and S in $H^s(\Sigma)$ given by

$$T\chi = \sum_{n\in\mathbb{Z}} \sum_{m\geq 0} \left(\chi, \chi_\mu^{(n,m)}\right)_\Sigma \chi_\nu^{(n,m)}, \tag{3.37}$$

$$S\chi = \sum_{n\in\mathbb{Z}} \sum_{m\geq 0} \left(\chi, \chi_{\bar\nu}^{(n,m)}\right)_\Sigma \chi_\mu^{(n,m)}. \tag{3.38}$$

(i) Let us first prove that these expressions actually define two bounded operators on $H^s(\Sigma)$. To see that T is bounded, consider the quantity $[T\chi - \chi]^2_{\mu,s}$, for $\chi \in H^s(\Sigma)$, which expands, by (3.16), in the form

$$\sum_{n'\in\mathbb{Z}} \sum_{m'\geq 0} (1+n'^2+m'^2)^s \left| \sum_{n\in\mathbb{Z}} \sum_{m\geq 0} (\chi, \chi_\mu^{(n,m)})_\Sigma \, (\chi_\nu^{(n,m)} - \chi_\mu^{(n,m)}, \chi_\mu^{(n',m')})_\Sigma \right|^2,$$

and simplifies to

$$\sum_{n\in\mathbb{Z}} \sum_{m'\geq 0} (1+n^2+m'^2)^s \left| \sum_{m\geq 0} (\chi, \chi_\mu^{(n,m)})_\Sigma \, (\tau_\nu^{(m)} - \tau_\mu^{(m)}, \tau_\mu^{(m')})_{\mathcal{I}} \right|^2,$$

by virtue of (3.36). Hence, by the Schwarz inequality, we have

$$[T\chi - \chi]^2_{\mu,s} \leq A_{\nu,s}\, [\chi]^2_{\mu,s}, \tag{3.39}$$

where $A_{\nu,s}$ is the bounded quantity defined in Lemma 3.5: the boundedness of T then follows from the triangle inequality. On the other hand, $[S\chi - \chi]^2_{\mu,s}$ expands as follows:

$$\sum_{n'\in\mathbb{Z}} \sum_{m'\geq 0} (1+n'^2+m'^2)^s \left| \left(\chi, \chi_{\bar\nu}^{(n',m')} - \chi_\mu^{(n',m')}\right)_\Sigma \right|^2.$$

Noticing that

$$\begin{aligned}\left(\chi, \chi_{\bar\nu}^{(n',m')} - \chi_\mu^{(n',m')}\right)_\Sigma &= \sum_{n\in\mathbb{Z}} \sum_{m\geq 0} \left(\chi, \chi_\mu^{(n,m)}\right)_\Sigma \left(\chi_\mu^{(n,m)}, \chi_{\bar\nu}^{(n',m')} - \chi_\mu^{(n',m')}\right)_\Sigma, \\ &= \sum_{m\geq 0} \left(\chi, \chi_\mu^{(n',m)}\right)_\Sigma \left(\tau_\mu^{(m)}, \tau_{\bar\nu}^{(m')} - \tau_\mu^{(m')}\right)_{\mathcal{I}},\end{aligned}$$

we deduce as above that $[S\chi - \chi]^2_{\mu,s} \leq B_{\nu,s}\,[\chi]^2_{\mu,s}$, where

$$B_{\nu,s} = \sup_{n\in\mathbb{Z}} \sum_{m\geq 0}\sum_{m'\geq 0} \left(\frac{1+n^2+m^2}{1+n^2+m'^2}\right)^s \left|(\tau_\mu^{(m')}, \tau_{\bar{\nu}}^{(m)} - \tau_\mu^{(m)})_{\mathcal{I}}\right|^2 .$$

From (3.28), we have $B_{\nu,s} = A_{\nu,-s}$. Lemma 3.5 thus shows that S is also bounded on $H^s(\Sigma)$.

(ii) Noticing that $T\chi_\mu^{(n,m)} = \chi_\nu^{(n,m)}$ and conversely $S\chi_\nu^{(n,m)} = \chi_\mu^{(n,m)}$, we infer that T is invertible and S is a left inverse of T, since $\mathcal{X}_\mu$ is an orthogonal basis of $H^s(\Sigma)$. To see that S is also a right inverse of T, it is sufficient to consider T and S as operators defined on $L^2(\Sigma)$ (for they both map $H^s(\Sigma)$ onto itself). In this case, this amounts to showing that the range $R(T)$ of T is the whole space $L^2(\Sigma)$. This latter property readily follows from Lemma 3.4. Indeed, $(\chi, \chi')_\Sigma = 0$ for all $\chi' \in R(T)$ implies, in particular, $(\chi, \chi_{\bar{\nu}}^{(n,m)})_\Sigma = 0$ for all $n \in \mathbb{Z}$ and $m \geq 0$, which shows that $\chi = 0$ since $\mathcal{X}_{\bar{\nu}}$ is a basis of $L^2(\Sigma)$.

As a consequence, T and S are bounded operators on $H^s(\Sigma)$ which are inverse to each other: this clearly shows that $\mathcal{X}_\nu$ is a basis of $H^s(\Sigma)$.

(iii) Finally, noticing that $[S\chi]_{\mu,s} = [\chi]_{\bar{\nu},s}$ (where $[\chi]_{\bar{\nu},s}$ is defined as in (3.16)), we infer that $[\chi]_{\nu,s}$ is a norm on $H^s(\Sigma)$ for every $\nu \in \mathbb{D}$. This completes the proof of Theorem 3.1. □

3.3. Analyticity of the coupling operator. We have seen that every root $\zeta_\nu^{(m)}$ of the dispersion equation (3.4) is analytic in $\mathbb{D}$, and consequently, every term of the series (3.23) is defined at any point of $\mathbb{D}$ provided that $K_n(\zeta_\nu^{(m)} r_0)$ does not vanish. This may happen only in the case $m = 0$: indeed, the set Z_n of zeros of the modified Bessel function K_n is contained in the region $|\arg z| > \pi/2$, and we know from (B.35) that $\zeta_\nu^{(0)}$ is the only root of (3.4) which is located in this domain (if $\nu \in \mathbb{D}^-$). Let $Z = \bigcup_{n\geq 0} Z_n$, and consider the set

$$\mathbb{K} = \{\nu = -z/r_0 \tan(z/r_0) \in \mathbb{C}^-;\ z \in Z\}. \tag{3.40}$$

Lemma 3.6. *For every $\nu \in \mathbb{D}^- \setminus \mathbb{K}$ and every $\chi \in H^{1/2}(\Sigma)$, the series* (3.23) *converges in $H^{-1/2}(\Sigma)$.*

Proof. Let $\nu \in \mathbb{D}^- \setminus \mathbb{K}$. Let us first study the asymptotic behaviour of $q_\nu^{(n,m)}$ when $n^2 + m^2 \to +\infty$. Using the uniform asymptotic expansions (A.20) and (A.21) of K_n, and noticing, from (3.33), that for large enough $n^2 + m^2$, the quantities $\zeta_\nu^{(m)} r_0/n$ lie in a closed subset of the validity domain of these expansions, we deduce, as in the proof of Lemma A.1,

$$\frac{K_n'(\zeta_\nu^{(m)} r_0)}{K_n(\zeta_\nu^{(m)} r_0)} = -\frac{\left(n^2 + \zeta_\nu^{(m)^2} r_0^2\right)^{1/2}}{\zeta_\nu^{(m)} r_0} \left\{1 + O\left(\left(n^2 + \zeta_\nu^{(m)^2} r_0^2\right)^{-1/2}\right)\right\}.$$

Consequently,

$$q_\nu^{(n,m)} = \left(r_0^{-2} n^2 + \zeta_\nu^{(m)^2}\right)^{1/2} + O(1). \tag{3.41}$$

It follows from the asymptotic behaviour (3.33) of $\zeta_\nu^{(m)}$ that

$$\left|q_\nu^{(n,m)}\right| \leq C\,(1 + n^2 + m^2)^{1/2}, \tag{3.42}$$

which shows that for every χ and χ' in $H^{1/2}(\Sigma)$,

$$\left|\langle \mathcal{Q}_\nu \chi, \chi' \rangle_{1/2,\Sigma}\right| = \left| \sum_{n\in\mathbb{Z}} \sum_{m\geq 0} q_\nu^{(n,m)} \, (\chi, \chi_{\bar\nu}^{(n,m)})_\Sigma \, (\chi_\nu^{(n,m)}, \chi')_\Sigma \right| \leq C \, [\chi]_{\bar\nu,1/2} \, [\chi']_{\nu,1/2},$$

by the Schwarz inequality (note that the expression of $\langle \mathcal{Q}_\nu \chi, \chi' \rangle_{1/2,\Sigma}$ results from (3.31)). The conclusion thus follows from Theorem 3.1. □

THEOREM 3.2. *The diagonal form* (3.23) *of the coupling operator* $\mathcal{Q}_\nu$ *defines an analytic family from* $\mathbb{D} \setminus \mathbb{K}$ *into the space* $\mathcal{B}(H^{1/2}(\Sigma), H^{-1/2}(\Sigma))$ *of bounded operators from* $H^{1/2}(\Sigma)$ *to* $H^{-1/2}(\Sigma)$.

Proof. Every term of the series (3.23) is obviously an analytic function of ν. We thus have to prove that the same holds for the sum of this series, which amounts to verifying that $\langle \mathcal{Q}_\nu \chi, \chi' \rangle_{1/2,\Sigma}$ is analytic in $\mathbb{D} \setminus \mathbb{K}$ for every χ and χ' in $H^{1/2}(\Sigma)$ (see Kato [13]). A sufficient condition for this property to be satisfied is, for instance,

$$\lim_{L\to\infty} \left(\sup_{\nu\in K} \sum_{n^2+m^2\geq L} q_\nu^{(n,m)} \, (\chi, \chi_{\bar\nu}^{(n,m)})_\Sigma \, (\chi_\nu^{(n,m)}, \chi')_\Sigma \right) = 0$$

for every compact set $K \subset \mathbb{D} \setminus \mathbb{K}$: this condition expresses that the sequence of partial sums of the series converges locally uniformly on $\mathbb{D} \setminus \mathbb{K}$ (see Henrici [7]). First, notice that the asymptotic expansion (3.33) of $\zeta_\nu^{(m)}$, as well as the estimates which derive from it, are valid uniformly with respect to $\nu \in K$. Consequently, by (3.42) and the Schwarz inequality, it is enough to show that for every $\chi \in H^{1/2}(\Sigma)$,

$$\lim_{L\to\infty} \left(\sup_{\nu\in K} \sum_{n^2+m^2\geq L} (1+n^2+m^2)^{1/2} \left| (\chi, \chi_{\bar\nu}^{(n,m)})_\Sigma \right|^2 \right) = 0.$$

Let P_L denote the orthogonal projection on the closed linear manifold of $H^{1/2}(\Sigma)$ spanned by $\{\chi_\mu^{(n,m)};\ n^2+m^2 \geq L\}$. We easily obtain

$$\sum_{n^2+m^2\geq L} \left| q_\nu^{(n,m)} \right| \left| (\chi, \chi_{\bar\nu}^{(n,m)})_\Sigma \right|^2 \leq C \, [P_L S \chi]_{\mu,1/2}^2 ,$$

where S is the bounded operator on $H^{1/2}(\Sigma)$ given by (3.38). We obviously have $[P_L S\chi]_{\mu,1/2} \to 0$ when $L \to +\infty$, locally uniformly with respect to ν. The statement of Theorem 3.2 follows. □

Remark 3.7. The analytic continuation of the roots $\zeta_\nu^{(m)}$ of the dispersion equation also provides the analytic continuation of the solution $\check{\mathcal{R}}_\nu \chi$ of the outer problem $\check{\mathcal{P}}_\nu$ (see Proposition 2.2). Indeed, if $\nu \in \mathbb{R}^+ \cup \mathbb{D}^+$, we have by construction

$$\check{\mathcal{R}}_\nu \chi = \sum_{n\in\mathbb{Z}} \sum_{m\geq 0} (\chi, \chi_{\bar\nu}^{(n,m)})_\Sigma \, \check\varphi_\nu^{(n,m)} \quad \forall \chi \in H^{1/2}(\Sigma). \tag{3.43}$$

By the same techniques as in §A.1, it may be easily seen that if $\nu \in \mathbb{D}^- \setminus \mathbb{K}$, this series converges in $H^1_{\mathrm{loc}}(\check\Omega)$. Moreover, it depends analytically on ν in $\mathbb{D} \setminus \mathbb{K}$ (same proof as for Theorem 3.2). This expansion thus defines an analytic family from $\mathbb{D} \setminus \mathbb{K}$ to $H^1_{\mathrm{loc}}(\check\Omega)$ (recall that $\check{\mathcal{R}}_\nu \chi \in H^1_{\mathrm{loc}}(\check\Omega)$ is analytic if, for every open bounded set $\mathcal{O} \subset \check\Omega$, the family $(\tilde{\mathcal{R}}_\nu\chi)_{|\mathcal{O}} \in H^1(\mathcal{O})$ is itself analytic). Note that $\tilde{\mathcal{R}}_\nu\chi$ becomes exponentially increasing at infinity when $\nu \in \mathbb{D}^- \setminus \mathbb{K}$, except if $(\chi, \chi_\nu^{(n,0)})_\Sigma = 0$ for all $n \in \mathbb{Z}$.

4. Eigenfrequencies and scattering frequencies.

In §2.4, we have defined problem $\hat{\mathcal{P}}_\nu$ (set on the bounded fluid domain $\hat{\Omega}$) by means of a coupling condition on the fictitious boundary Σ. Theorem 3.2 shows that the coupling operator $\mathcal{Q}_\nu$ actually is an analytic family in $\mathbb{D} \setminus \mathbb{K}$: consequently, we are now able to extend $\hat{\mathcal{P}}_\nu$ to $\mathbb{D}^- \setminus \mathbb{K}$.

4.1. Analytic continuation of the reduced problem.

Let $\nu \in \mathbb{D} \setminus \mathbb{K}$; a variational formulation of problem $\hat{\mathcal{P}}_\nu$ reads (see (2.30)):

$$\text{(4.1)}\qquad \begin{array}{l} \text{Find } \hat{X} = (\varphi, \hat{u}) \in \hat{\mathcal{H}} \text{ such that} \\ \hat{a}_\nu(\hat{X}, \hat{Y}) = \hat{l}(f; \hat{Y}) \quad \forall \hat{Y} \in \hat{\mathcal{H}}, \end{array}$$

where $\hat{a}_\nu(\cdot\,,\cdot)$ is the sesquilinear form defined on $\hat{\mathcal{H}} \times \hat{\mathcal{H}}$ by

$$\text{(4.2)}\qquad \hat{a}_\nu(\hat{X}, \hat{Y}) = \hat{k}(\hat{X}, \hat{Y}) - \nu\,\hat{m}(\hat{X}, \hat{Y}) - \nu^{1/2}\,\hat{c}(\hat{X}, \hat{Y}) + \hat{q}_\nu(\hat{X}, \hat{Y}),$$

with

$$\text{(4.3)}\quad \hat{k}(\hat{X}, \hat{Y}) = \int_{\hat{\Omega}} \nabla\varphi.\overline{\nabla\psi}\, d\hat{\Omega} + \int_B a_{ijkh} e_{ij}(\hat{u}) \overline{e_{kh}(\hat{v})}\, dB,$$

$$\text{(4.4)}\quad \hat{m}(\hat{X}, \hat{Y}) = \int_{\hat{S}} \varphi\overline{\psi}\, d\hat{S} + \int_B \rho\, \hat{u}.\overline{\hat{v}}\, dB,$$

$$\text{(4.5)}\quad \hat{c}(\hat{X}, \hat{Y}) = \int_\Gamma (\hat{u}.n\,\overline{\psi} + \varphi\,\overline{\hat{v}.n})\, d\Gamma,$$

$$\text{(4.6)}\quad \hat{q}_\nu(\hat{X}, \hat{Y}) = <\mathcal{Q}_\nu\varphi, \psi>_{1/2,\Sigma} = \sum_{n\in\mathbb{Z}} \sum_{m\geq 0} q_\nu^{(n,m)} \left(\varphi, \chi_\nu^{(n,m)}\right)_\Sigma \left(\chi_\nu^{(n,m)}, \psi\right)_\Sigma$$

for all $\hat{X} = (\varphi, \hat{u})$ and $\hat{Y} = (\psi, \hat{v})$ in $\hat{\mathcal{H}}$. The semilinear form $\hat{l}$, which depends on the datum $f = (f', f'') \in \mathcal{F}$, is given by

$$\text{(4.7)}\qquad \hat{l}(f; \hat{Y}) = -\int_\Gamma (f'\overline{\psi} + f''\overline{\hat{v}.n})\, d\Gamma.$$

Let $\hat{A}_\nu$ and $\hat{L}(f)$ be, respectively, the bounded operator on $\hat{\mathcal{H}}$ and the element of $\hat{\mathcal{H}}$ associated with $\hat{a}_\nu(\cdot\,,\cdot)$ and $\hat{l}(f;\cdot)$, i.e.,

$$\text{(4.8)}\qquad (\hat{A}_\nu\hat{X}, \hat{Y})_{\hat{\mathcal{H}}} = \hat{a}_\nu(\hat{X}, \hat{Y}) \quad\text{and}\quad (\hat{L}(f), \hat{Y})_{\hat{\mathcal{H}}} = \hat{l}(f; \hat{Y}).$$

The variational formulation (4.1) of $\hat{\mathcal{P}}_\nu$ then amounts to $\hat{A}_\nu\hat{X} = \hat{L}(f)$. The object of this paragraph is to prove the following statement.

THEOREM 4.1. *The operators $\hat{A}_\nu$ for $\nu \in \mathbb{D} \setminus \mathbb{K}$ form a holomorphic family of Fredholm operators on $\hat{\mathcal{H}}$. Moreover, $\hat{A}_\nu^{-1}$ is a meromorphic family in $\mathbb{D} \setminus \mathbb{K}$ whose poles are located in $\mathbb{R}^+$ or $\mathbb{D}^- \setminus \mathbb{K}$.*

Remark 4.1. The poles of $\hat{A}_\nu^{-1}$ are the values of ν for which the Fredholm operator $\hat{A}_\nu$ is not invertible. In other words, they are the solutions of the following nonlinear (and nonselfadjoint) eigenvalue problem:

$$\text{(4.9)}\qquad \begin{array}{l} \text{Find } \nu \in \mathbb{R}^+ \cup \mathbb{D}^- \setminus \mathbb{K} \text{ such that there exists } \hat{X} \in \hat{\mathcal{H}} \setminus \{0\} \text{ which satisfies} \\ \hat{k}(\hat{X}, \hat{Y}) - \nu^{1/2}\,\hat{c}(\hat{X}, \hat{Y}) + \hat{q}_\nu(\hat{X}, \hat{Y}) = \nu\,\hat{m}(\hat{X}, \hat{Y}) \quad \forall \hat{Y} \in \hat{\mathcal{H}}. \end{array}$$

Proof. (i) The analyticity of the family $\hat{A}_\nu$ follows readily from Theorem 3.2.

(ii) Let us prove that in every compact subset K of $\mathbb{D}\setminus\mathbb{K}$, $\hat{A}_\nu$ can be written as

$$\hat{A}_\nu = \hat{J}_\nu + \hat{K}_\nu,$$

where $\hat{J}_\nu$ and $\hat{K}_\nu$ are two analytic families of bounded operators on $\hat{\mathcal{H}}$, $\hat{J}_\nu$ is invertible with bounded inverse, and $\hat{K}_\nu$ is compact. The main difficulty arises from the coupling term $\hat{q}_\nu$ which will be rewritten as the sum of two forms such that the real part of the first one is positive, and the second one corresponds to a compact operator on $\hat{\mathcal{H}}$.

Consider the bounded operators $\hat{J}_\nu$ and $\hat{K}_\nu$ defined on $\hat{\mathcal{H}}$ by

$$\begin{aligned}(\hat{J}_\nu\hat{X},\hat{Y})_{\hat{\mathcal{H}}} &= \hat{k}(\hat{X},\hat{Y}) + \hat{m}(\hat{X},\hat{Y}) + \hat{q}_{d,\nu}(\hat{X},\hat{Y}),\\ (\hat{K}_\nu\hat{X},\hat{Y})_{\hat{\mathcal{H}}} &= -(\nu+1)\,\hat{m}(\hat{X},\hat{Y}) - \nu^{1/2}\hat{c}(\hat{X},\hat{Y}) + \hat{q}_\nu(\hat{X},\hat{Y}) - \hat{q}_{d,\nu}(\hat{X},\hat{Y}),\end{aligned}$$

where $\hat{q}_{d,\nu}$ is the sesquilinear form given by

$$(4.10)\qquad \hat{q}_{d,\nu}(\hat{X},\hat{Y}) = \sum_{n\in\mathbb{Z}}\sum_{m\geq 0}\left(d + r_0^{-2}n^2 + \zeta_\nu^{(m)^2}\right)^{1/2} (\hat{\varphi},\chi_\nu^{(n,m)})_\Sigma\,(\chi_\nu^{(n,m)},\hat{\psi})_\Sigma,$$

where d is a real positive parameter. We prove in Lemma 4.1 below that for large enough d, $\hat{q}_{d,\nu}(\hat{X},\hat{X})$ depends analytically on ν in K, and its real part is nonnegative for every $\hat{X}\in\hat{\mathcal{H}}$. Moreover, as in the proof of Proposition 2.1, it may be easily seen that $\hat{k}+\hat{m}$ is a continuous coercive symmetric form on $\hat{\mathcal{H}}\times\hat{\mathcal{H}}$. Consequently, $\hat{J}_\nu$ defines a holomorphic family of bounded invertible operators with bounded inverses, and the inverse family $\hat{J}_\nu^{-1}$ is also holomorphic in K (see Kato [13, §VII-1.1]).

On the other hand, the compactness of $\hat{K}_\nu$ is a straightforward consequence of Lemma 4.2 below and the Rellich theorem.

(iii) Rewriting $\hat{A}_\nu^{-1}$ in the form $(1+\hat{J}_\nu^{-1}\hat{K}_\nu)^{-1}\,\hat{J}_\nu^{-1}$, we can use Steinberg's theorem [31] which states that either $(1+\hat{J}_\nu^{-1}\hat{K}_\nu)$ is nowhere invertible in K, or else $(1+\hat{J}_\nu^{-1}\hat{K}_\nu)^{-1}$ is meromorphic in K. We are of course in the second case, since problem $\hat{\mathcal{P}}_\nu$ is well posed if $\nu\in\mathbb{D}^+$ (see Remark 2.4). Consequently, $\hat{A}_\nu^{-1}$ is a meromorphic family which has no poles in $\mathbb{D}^+$. □

LEMMA 4.1. *For every compact subset K of $\mathbb{D}\setminus\mathbb{K}$, there exists $d>0$ such that*

$$(4.11)\qquad \operatorname{Re}\left(\hat{q}_{d,\nu}(\hat{X},\hat{X})\right)\geq 0\quad \forall\hat{X}\in\hat{\mathcal{H}}\quad\forall\nu\in K.$$

Moreover, $\hat{q}_{d,\nu}(\hat{X},\hat{X})$ depends analytically on ν in K.

Proof. Consider the sesquilinear form $t_{d,\nu}$ introduced in the proof of Proposition 3.1:

$$t_{d,\nu}(\chi,\chi') = \int_\Sigma \nabla_\sigma\chi..\overline{\nabla_\sigma\chi'}\,d\Sigma - \nu\int_{\mathcal{C}_0}\chi\overline{\chi'}\,d\mathcal{C} + d\int_\Sigma\chi\overline{\chi'}\,d\Sigma\quad\forall(\chi,\chi')\in H^1(\Sigma)^2.$$

We know from Lions's lemma that we can choose $d>0$ such that $t_{d,\nu}$ is a strictly m-accretive form for every $\nu\in K$. Consequently, there exists a family of strictly m-accretive unbounded operators $T_{d,\nu}$ on $L^2(\Sigma)$ given by

$$(T_{d,\nu}\chi,\chi')_\Sigma = t_{d,\nu}(\chi,\chi')\quad\forall\chi\in D(T_{d,\nu})\quad\forall\chi'\in H^1(\Sigma).$$

Let $T_{d,\nu}^{1/2}$ be the unique strictly m-accretive operator such that $(T_{d,\nu}^{1/2})^2 = T_{d,\nu}$ (see Kato [13, Thm. V-3.35]). In particular, we have

$$\operatorname{Re}(T_{d,\nu}^{1/2}\chi,\chi)_\Sigma \geq 0 \quad \forall \chi \in D(T_{d,\nu}^{1/2}).$$

We prove below that $D(T_{d,\nu}^{1/2}) = H^1(\Sigma)$ and $T_{d,\nu}^{1/2}$ expands as follows:

$$T_{d,\nu}^{1/2}\chi = \sum_{n\in\mathbb{Z}}\sum_{m\geq 0} \left(d + r_0^{-2}n^2 + \zeta_\nu^{(m)^2}\right)^{1/2} (\chi, \chi_{\bar\nu}^{(n,m)})_\Sigma\, \chi_\nu^{(n,m)}, \tag{4.12}$$

which of course implies (4.11), since $H^1(\Sigma)$ is dense in $H^{1/2}(\Sigma)$.

First, notice that the domain $H^1(\Sigma)$ of the form $t_{d,\nu}$ does not depend on ν, and $t_{d,\nu}(\chi,\chi)$ is holomorphic in K for each fixed $\chi \in H^1(\Sigma)$: the operators $T_{d,\nu}$ associated with such a family of forms (which is called holomorphic of type (a)) define a holomorphic family (Kato [13, Thm. VII-4.2]). Consequently, the family $T_{d,\nu}^{1/2}$ is also holomorphic in K (Kato [13, Rem. VII-4.7]).

On the other hand, we have seen in the proof of Proposition 3.1 that for real positive ν, the expansion (4.12) of $T_{d,\nu}^{1/2}\chi$ is valid for every $\chi \in D(T_{d,\nu}^{1/2}) = H^1(\Sigma)$. Moreover, using exactly the same techniques as for Theorem 3.2, we easily prove that the series in (4.12) defines a holomorphic family of unbounded operators on $L^2(\Sigma)$ (with domain $H^1(\Sigma)$). By virtue of the unique continuation property (Kato [13, Rem. VII-1.6]), this family is nothing but $T_{d,\nu}^{1/2}$, which completes the proof. □

LEMMA 4.2. *Let K be a compact subset of $\mathbb{D}\setminus\mathbb{K}$ and $d > 0$; then, for every $\nu \in K$, we have*

$$\left|\hat q_\nu(\hat X,\hat Y) - \hat q_{d,\nu}(\hat X,\hat Y)\right| \leq C\, \|\hat\varphi\|_\Sigma \left\|\hat\psi\right\|_\Sigma, \tag{4.13}$$

for all $\hat X = (\hat\varphi,\hat u)$ and $\hat Y = (\hat\psi,\hat v)$ in $\hat{\mathcal{H}}$.

Proof. We deduce from the asymptotic behaviour (3.41) of $q_\nu^{(n,m)}$ that

$$\left|\hat q_\nu(\hat X,\hat Y) - \hat q_{d,\nu}(\hat X,\hat Y)\right| \leq C\sum_{n\in\mathbb{Z}}\sum_{m\geq 0}\left|(\hat\varphi,\chi_{\bar\nu}^{(n,m)})_\Sigma\,(\chi_\nu^{(n,m)},\hat\psi)_\Sigma\right|.$$

Relation (4.13) thus results from Theorem 3.1 by the Schwarz inequality. □

4.2. Singularities of the reduced problem. The purpose of this paragraph is to study how the poles of $\hat A_\nu^{-1}$ affect the solution of the reduced problem. Indeed, it may happen that for a datum $f \in \mathcal{F}$, the solution $\hat A_\nu^{-1}\hat L(f)$ of $\hat{\mathcal{P}}_\nu$ is regular in a vicinity of a pole ν_* of $\hat A_\nu^{-1}$: this situation may occur if the datum f is such that $\hat L(f)$ belongs to the range $\mathrm{R}(\hat A_{\nu_*})$ of $\hat A_{\nu_*}$. The question is whether there exists some datum such that the solution of $\hat{\mathcal{P}}_\nu$ becomes singular in the vicinity of ν_*. We first prove the following result which concerns the nonreal poles of $\hat A_\nu^{-1}$.

PROPOSITION 4.1. *Let $\nu_* \in \mathbb{D}^-\setminus\mathbb{K}$. Then, ν_* is a pole of $\hat A_\nu^{-1}$ if and only if there exists at least a datum $f \in \mathcal{F}$ such that ν_* is a pole of $\hat A_\nu^{-1}\hat L(f)$.*

Proof. (i) If ν_* is a pole of $\hat A_\nu^{-1}\hat L(f)$ for some $f \in \mathcal{F}$, it is clearly a pole of $\hat A_\nu^{-1}$: a singularity of the solution can only proceed from a singularity of $\hat A_\nu^{-1}$.

(ii) Conversely, assume that ν_* is a pole of $\hat A_\nu^{-1}$. Let us prove by contradiction that ν_* is a pole of $\hat A_\nu^{-1}\hat L(f)$ for some $f \in \mathcal{F}$. Suppose that

$$\hat L(f) \in \mathrm{R}(\hat A_{\nu_*}) = \left(\operatorname{Ker}(\hat A_{\nu_*}^*)\right)^\perp \quad \forall f \in \mathcal{F}.$$

Let $\hat{X} = (\hat{\varphi}, \hat{u}) \in \operatorname{Ker}(\hat{A}^*_{\nu_*})$. From the definition (4.7), (4.8) of $\hat{L}(f)$, the relation $(\hat{L}(f), \hat{X})_{\hat{\mathcal{H}}} = 0$ yields

$$\int_\Gamma (f' \overline{\hat{\varphi}} + f'' \overline{\hat{u}.n})\, d\Gamma = 0 \quad \forall f = (f', f'') \in \mathcal{F},$$

which shows that

$$\hat{\varphi} = 0 \text{ and } \hat{u}.n = 0 \quad \text{on } \Gamma. \tag{4.14}$$

On the other hand, it may be easily seen that $\hat{A}^*_{\nu_*}$ is given by

$$(\hat{A}^*_{\nu_*} \hat{X}, \hat{Y})_{\hat{\mathcal{H}}} = \hat{k}(\hat{X}, \hat{Y}) - \overline{\nu_*}\, \hat{m}(\hat{X}, \hat{Y}) - \overline{\nu_*^{1/2}}\, \hat{c}(\hat{X}, \hat{Y}) + \overline{\hat{q}_\nu(\hat{Y}, \hat{X})}.$$

As $\hat{X}$ is assumed to belong to $\operatorname{Ker}(\hat{A}^*_{\nu_*})$ and $\hat{c}(\hat{X}, \hat{Y}) = 0$ by virtue of (4.14), we infer that

$$\hat{k}(\hat{X}, \hat{Y}) - \overline{\nu_*}\, \hat{m}(\hat{X}, \hat{Y}) + \overline{\hat{q}_\nu(\hat{Y}, \hat{X})} = 0 \quad \forall \hat{Y} \in \hat{\mathcal{H}}.$$

This implies in particular that $\hat{\varphi}$ and $\hat{u}$ satisfy, respectively, the equations

$$\begin{aligned} &\Delta\hat{\varphi} = 0 \quad \text{in } \hat{\Omega}, \\ &\partial_n \hat{\varphi} = 0 \quad \text{on } \Gamma, \end{aligned} \tag{4.15}$$

and

$$\begin{aligned} &\overline{\nu_*} \rho\, \hat{u}_i + \partial_{x_j} \sigma_{ij}(\hat{u}) = 0 \quad \text{in } B, \\ &\sigma_{ij}(\hat{u})\, n_j = 0 \quad \text{on } \Gamma_0 \cup \Gamma. \end{aligned} \tag{4.16}$$

Equations (4.15) together with (4.14) show that $\hat{\varphi} = 0$ in $\hat{\Omega}$. From (4.16), we readily deduce that $\hat{u} = 0$ in B, since $\operatorname{Im} \nu_* \neq 0$. As a consequence, $\operatorname{Ker}(\hat{A}^*_{\nu_*}) = \{0\}$, which is inconsistent with the fact that ν_* is a pole of $\hat{A}_\nu^{-1}$. □

Assume now that ν_* is a real positive pole of $\hat{A}_\nu^{-1}$. In this case, the arguments of the proof of Proposition 4.1 are still valid, except that (4.16) does not imply $\hat{u} = 0$ in B. Indeed, (4.16) is a classical eigenvalue problem which has a countable infinity of real positive eigenvalues, and ν_* may be one of them. In fact, from (4.14), the problem to be dealt with is the following:

$$\begin{aligned} &\nu_* \rho\, \hat{u}_i + \partial_{x_j} \sigma_{ij}(\hat{u}) = 0 \quad \text{in } B, \\ &\sigma_{ij}(\hat{u})\, n_j = 0 \quad \text{on } \Gamma_0 \cup \Gamma, \\ &\hat{u}.n = 0 \quad \text{on } \Gamma. \end{aligned} \tag{4.17}$$

A real positive number ν_* such that (4.17) has a nonzero solution $\hat{u}$ is referred to as an exceptional eigenvalue. Hargé [4] has proved that for almost every elastic body B with regular boundary ∂B, there is no exceptional eigenvalue. This means that these values may exist only for very particular shapes of the body. The axisymmetrical bodies provide the simplest example of exceptional eigenvalues. Indeed, the revolution motions about the axis of symmetry are obviously solutions of problem (4.17) associated with the value $\nu_* = 0$. But even in this case, it is not clear whether there exists a nonzero exceptional eigenvalue.

An exceptional eigenvalue ν_* is always a pole of $\hat{A}_\nu^{-1}$, since we can find $\hat{u} \neq 0$ such that $\hat{A}_\nu(0, \hat{u}) = 0$. Nevertheless, we cannot assert that there exists a datum f such that it is a pole of $\hat{A}_\nu^{-1} \hat{L}(f)$: indeed, if $\operatorname{Ker}(\hat{A}^*_{\nu_*})$ reduces to the pairs $(0, \hat{u})$ where $\hat{u}$

is a solution of (4.17), then $\hat{L}(f) \in R(\hat{A}_{\nu_*})$ for all $f \in \mathcal{F}$. As a consequence, for real poles of $\hat{A}_\nu^{-1}$, the statement of Proposition 4.1 becomes the following.

PROPOSITION 4.2. *Let $\nu_* \in \mathbb{R}^+$. Then, ν_* is a pole of $\hat{A}_\nu^{-1}$ if and only if there exists at least a datum $f \in \mathcal{F}$ such that ν_* is a pole of $\hat{A}_\nu^{-1}\hat{L}(f)$, except maybe when ν_* is an exceptional eigenvalue of the body.*

4.3. Analytic continuation of the initial problem. We are now able to proceed to the extension to the lower complex half plane of the solution operator $\mathcal{R}_\nu$ associated with the initial problem $\mathcal{P}_\nu$. This operator was defined in §§2.3 and 2.4 for $\nu \in \mathbb{R}^+ \cup \mathbb{C}^+$, and the equivalence between $\mathcal{P}_\nu$ and the reduced problem $\hat{\mathcal{P}}_\nu$ (Proposition 2.3) provides in this case a convenient expression for $\mathcal{R}_\nu$. Indeed, for a given $f \in \mathcal{F}$, let $\hat{X} = (\hat{\varphi}, \hat{u})$ denote the solution of $\hat{\mathcal{P}}_\nu$ (i.e., $\hat{X} = \hat{A}_\nu^{-1}\hat{L}(f)$). We know from Proposition 2.3 that the solution of $\mathcal{P}_\nu$ for the same datum f reads

$$\mathcal{R}_\nu f = (\varphi, u), \quad \text{where } (\varphi_{|\hat{\Omega}}, u) = (\hat{\varphi}, \hat{u}) \text{ and } \varphi_{|\check{\Omega}} = \check{\mathcal{R}}_\nu \hat{\varphi}_{|\Sigma}. \tag{4.18}$$

This identity actually defines the analytic continuation of $\mathcal{R}_\nu f$ to $\mathbb{D} \setminus \mathbb{K}$, since $\hat{A}_\nu^{-1}$ is a meromorphic family in $\mathbb{D} \setminus \mathbb{K}$ (Theorem 4.1) and $\check{\mathcal{R}}_\nu \hat{\varphi}_{|\Sigma}$ is analytic in this domain (Remark 3.7). As a consequence, for every $f \in \mathcal{F}$, the family $\nu \to \mathcal{R}_\nu f \in \mathcal{H}_{\text{loc}}$ is meromorphic in $\mathbb{D} \setminus \mathbb{K}$ and its poles coincide with those of $\hat{A}_\nu^{-1} f$. By virtue of Propositions 4.1 and 4.2, we can then state the main result of this paper.

THEOREM 4.2. *The solution operators $\mathcal{R}_\nu$, from $\mathcal{F}$ onto $\mathcal{H}_{\text{loc}}$, form a meromorphic family on $\mathbb{D} \setminus \mathbb{K}$. Its poles coincide with those of $\hat{A}_\nu^{-1}$, i.e., the solutions of the nonlinear eigenvalue problem* (4.9), *except the exceptional eigenvalues defined by* (4.17) *which may be removable singularities of $\mathcal{R}_\nu$. These poles are located in $\mathbb{R}^+$ or $\mathbb{D}^- \setminus \mathbb{K}$, and are referred to as "eigenfrequencies" in the first case, and "scattering frequencies" in the second one.*

Recall that the family $\mathcal{R}_\nu$ is said to be meromorphic near $\nu_* \in \mathbb{C}$ if there exists a vicinity $\mathcal{V}$ of ν_* and a sequence $\{\mathcal{R}^{(p)};\ p \geq -P\}$ (with $P \geq 0$) of operators from $\mathcal{F}$ to $\mathcal{H}_{\text{loc}}$ such that

$$\mathcal{R}_\nu f = \sum_{p \geq -P} (\nu - \nu_*)^p \mathcal{R}^{(p)} f, \tag{4.19}$$

where the series converges in $\mathcal{H}_{\text{loc}}$ for every $f \in \mathcal{F}$ and every $\nu \in \mathcal{V} \setminus \{\nu_*\}$. By (4.18), the $\mathcal{R}^{(p)}$ may be easily obtained from the series expansions of $\hat{A}_\nu^{-1}$ and $\check{\mathcal{R}}_\nu$.

Remark 4.2. The distinction between the real positive poles of $\mathcal{R}_\nu$ and the complex ones seems artificial, since they both are solutions of the same nonlinear eigenvalue problem. This difference actually results from the associated eigenvectors. Let $(\hat{\varphi}, \hat{u})$ be an eigenvector associated with a solution ν_* of problem (4.19), and let φ denote the analytic continuation of $\hat{\varphi}$ in the whole domain Ω, i.e., $\varphi_{|\hat{\Omega}} = \hat{\varphi}$ and $\varphi_{|\breve{\varphi}} = \check{\mathcal{R}}_{\nu_*} \hat{\varphi}_{|\Sigma}$. The asymptotic behaviour of φ at infinity depends on the location of ν_*: if ν_* is real, $\check{\mathcal{R}}_{\nu_*} \hat{\varphi}_{|\Sigma} \to 0$ at infinity (see §3.1), whereas it increases exponentially if $\nu_* \in \mathbb{D}^- \setminus \mathbb{K}$ (see Remark 3.7).

Remark 4.3. The points of $\mathbb{K}$ (defined by (3.40)), which are poles of the coupling operator, were left out for the analytic continuation of the reduced problem. In fact, these singularities are not intrinsic quantities of the initial problem, since they depend on the choice of the fictitious boundary Σ. This means that $\mathcal{R}_\nu$ actually defines a meromorphic family in $\mathbb{D}$. From a practical point of view, it is necessary to decide whether a point of $\mathbb{K}$ is a scattering frequency or not: a convenient criterion is given in [18, §4.3].

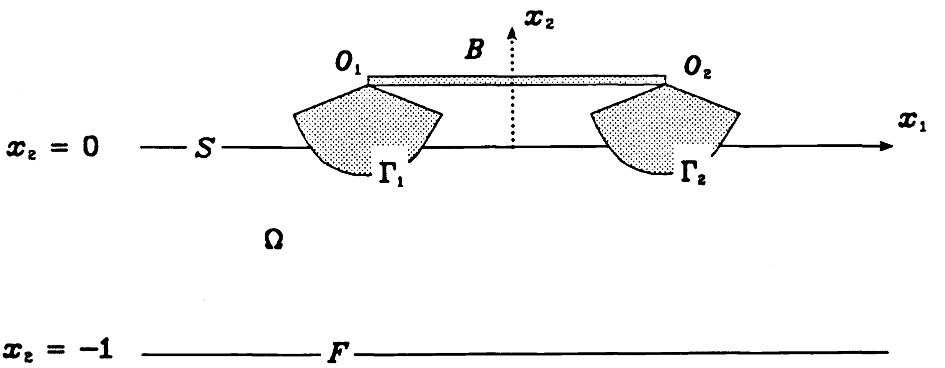


FIG. 5.1

Remark 4.4. On the other hand, the branch points of the solutions of the dispersion equation, which led us to introduce cuts in the complex plane, are clearly intrinsic singularities of the problem, for they do not depend on the method we used to construct the analytic continuation of $\mathcal{R}_\nu$. These are characteristic quantities of the asymptotic behaviour of the system at infinity.

Remark 4.5. The connection between scattering frequencies and resonant states (i.e., the peaks of the response curve of the system for real frequencies) has been circumstantially studied in [18]: for an internal approximation of the problem, we show how to construct explicitly the first terms of the Laurent series of the solution of $\hat{\mathcal{P}}_\nu$ in the vicinity of a scattering frequency. The same results apply to the present problem. A concrete example is given in §5.5 below.

5. Numerical application. We present in this last paragraph the outcome of a numerical application of our method for a two-dimensional problem. We are concerned here with the small motions of an elastic "catamaran" constituted of two rigid floating bodies linked together by an elastic beam. The characterization of the scattering frequencies associated with this problem is performed along the same lines as in the preceding sections: we will restrict ourselves to mention the basic stages of the method, as well as the main differences from the three-dimensional case (a more detailed study may be found in [5]). Except where otherwise stated, the notation previously introduced are retained.

5.1. Equations of the stationary problem. Consider a floating elastic "catamaran" as shown in Fig. 5.1. Let $\Omega \subset \mathbb{R}^2$ denote the domain filled by the fluid at rest, and $\partial\Omega$ its boundary which consists of the free surface S (located at $x_2 = 0$), the bottom F ($x_2 = -1$) and the immersed parts Γ_k, $k = 1, 2$, of the two hulls. The elastic beam B, which is parallel to S, is embedded in the bodies B_k at points O_k, $k = 1, 2$. For the sake of simplicity, it is assumed unstrained when the system is at rest. We denote by φ the (complex) velocity potential of the fluid, $u = (u_1, u_2)$ the displacement field of the beam (i.e., the horizontal and vertical displacements of each point of the beam), and s_k, $k = 1, 2$, the 3-components vectors which characterize the motions of the bodies with respect to O_k (displacement of O_k and rotation of the bodies):

$$s_k = (u_1(O_k), u_2(O_k), d_{x_1}u_2(O_k)). \tag{5.1}$$

The time-harmonic vibrations of the system are described by the following set of equations (where ν still denotes the frequency squared):

$$\Delta\varphi = 0 \quad \text{in } \Omega, \tag{5.2}$$

$$\partial_n\varphi - \nu\,\varphi = 0 \quad \text{on } S, \tag{5.3}$$

$$\partial_n\varphi = 0 \quad \text{on } F, \tag{5.4}$$

$$\nu\rho\, u_1 + d_{x_1}(\alpha\, d_{x_1} u_1) = 0 \quad \text{on } B, \tag{5.5}$$

$$\nu\rho\, u_2 - d^2_{x_1}(\beta\, d^2_{x_1} u_2) = 0 \quad \text{on } B, \tag{5.6}$$

$$\nu^{1/2}\, s_k.n_g - \partial_n\varphi = f'_k \quad \text{on } \Gamma_k, \quad k = 1,2, \tag{5.7}$$

$$(\nu\,\mathbb{M}_k - \mathbb{K}_k)s_k + \nu^{1/2}\int_{\Gamma_k} \varphi\, n_g\, d\Gamma_k + g_k = f''_k, \qquad k = 1,2. \tag{5.8}$$

Equations (5.2)–(5.4) are the same as in problem (2.13). Relations (5.5) and (5.6) are the dynamic equations of an elastic beam which model, respectively, the longitudinal and transverse vibrations of B. The mechanical data ρ, α, and β are assumed to belong to $L^\infty(B)$ and to be bounded from below by a positive constant (ρ is the mass of the beam per unit length, $\alpha = EA$ where E is the Young's modulus and A is the measure of each section of the beam, $\beta = EI$ where I is the geometrical inertia of each section). Equation (5.7) stands for the continuity of the normal velocity on the hulls: n_g denotes the so-called "generalized normal" defined by

$$n_g = (n_1, n_2, x_1 n_2 - x_2 n_1),$$

where $n = (n_1, n_2)$ is the unitary normal to Γ_k at point (x_1, x_2) measured with respect to O_k. Finally, (5.8) expresses the dynamic equations of the two rigid bodies (see John [10]). $\mathbb{M}_k$, which is referred to as the "generalized mass matrix," is the 3×3 positive matrix given by

$$\mathbb{M}_k = \begin{pmatrix} m_k & 0 & -m_k x_{2G_k} \\ 0 & m_k & m_k x_{1G_k} \\ -m_k x_{2G_k} & m_k x_{1G_k} & I_k \end{pmatrix},$$

where m_k is the mass of the body, (x_{1G_k}, x_{2G_k}) is the location of its centre of gravity with respect to O_k, and I_k is its inertia at point O_k. The "hydrostatic stiffness matrix" $\mathbb{K}_k$ is defined by

$$\mathbb{K}_k = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \int_{S_k} dS_k & \int_{S_k} x_1\, dS_k \\ 0 & \int_{S_k} x_1\, dS_k & (x_{2C_k} - x_{2G_k})\int_{\Omega_k} d\Omega_k + \int_{S_k} x_1^2\, dS_k \end{pmatrix},$$

where S_k and Ω_k are represented in Fig. 5.2, and C_k is the geometrical centre of gravity of Ω_k. The 3-components vector g_k is the torque of the force exerted on the body by the beam, which expresses as follows:

$$g_k = (-1)^k\,(-\alpha\, d_{x_1} u_1(O_k), d_{x_1}(\beta\, d^2_{x_1} u_2(O_k), -\beta\, d^2_{x_1} u_2(O_k))). \tag{5.9}$$

In Equations (5.7) and (5.8), the right-hand members depend on the incident wave: f'_k is a function defined on Γ_k, and $f''_k \in \mathbb{C}^3$. More precisely, if φ_I denotes the velocity potential of the incident wave, we have

$$f'_k = \partial_n\varphi_I \quad \text{and} \quad f''_k = -\nu^{1/2}\int_{\Gamma_k} \varphi_I\, n_g\, d\Gamma_k.$$

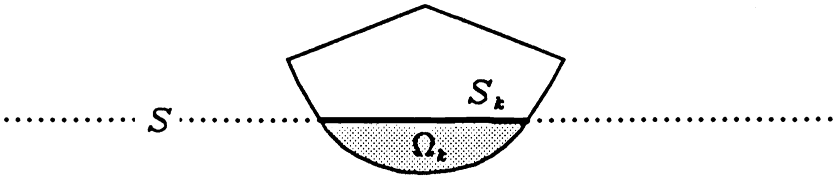


FIG. 5.2

As in §2.3, the asymptotic behaviour at infinity of the scattered wave is specified by means of the outgoing radiation condition

$$\lim_{x_1 \to \pm\infty} \int_{-1}^{0} \left|\partial_{x_1}\varphi \mp i\nu_0\,\varphi\right|^2 dx_2 = 0, \tag{5.10}$$

where ν_0 is, here again, the only positive root of

$$\nu_0 \tanh \nu_0 = \nu. \tag{5.11}$$

For every $\nu \in \mathbb{R}^+$, the stationary problem, denoted by $\mathcal{P}_\nu$ in the sequel, is then defined as follows:

$$\begin{array}{l} \text{Find } X = (\varphi, u) \in \mathcal{H}_{\text{loc}} \text{ such that} \\ \varphi \text{ and } u \text{ satisfy (5.2)–(5.8),} \\ \varphi \text{ satisfies the radiation condition (5.10),} \end{array} \tag{5.12}$$

where $\mathcal{H}_{\text{loc}}$ denotes the Frechet space

$$\mathcal{H}_{\text{loc}} = H^1_{\text{loc}}(\Omega) \times (H^1(B) \times H^2(B)), \tag{5.13}$$

and the datum $f = ((f_1', f_1''), (f_2', f_2''))$ is assumed to belong to

$$\mathcal{F} = (L^2(\Gamma_1) \times \mathbb{C}^3) \times (L^2(\Gamma_2) \times \mathbb{C}^3). \tag{5.14}$$

Following the same idea as in §2.4, problem $\mathcal{P}_\nu$ can be extended to the upper complex half plane by replacing the radiation condition by a decay condition at infinity. For every $\nu \in \mathbb{C}^+$, consider the problem (also denoted by $\mathcal{P}_\nu$)

$$\begin{array}{l} \text{Find } X = (\varphi, u) \in \mathcal{H} \text{ such that} \\ \varphi \text{ and } u \text{ satisfy (5.2)–(5.8),} \end{array} \tag{5.15}$$

where $\mathcal{H}$ is the Hilbert space

$$\mathcal{H} = H^1(\Omega) \times (H^1(B) \times H^2(B)). \tag{5.16}$$

PROPOSITION 5.1. *Let $\nu \in \mathbb{C}^+$. For every $f \in \mathcal{F}$, problem $\mathcal{P}_\nu$ has a unique solution $X = \mathcal{R}_\nu f$ in $\mathcal{H}$. Moreover, $\mathcal{R}_\nu$ is a continuous operator from $\mathcal{F}$ to $\mathcal{H}$.*

Proof. A variational formulation of $\mathcal{P}_\nu$ reads

$$\begin{array}{l} \text{Find } X \in \mathcal{H} \text{ such that} \\ a_\nu(X, Y) = l(f; Y) \quad \forall Y \in \mathcal{H}, \end{array}$$

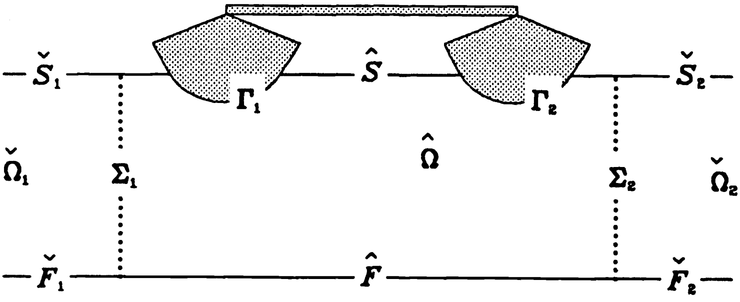


FIG. 5.3. *The inner and outer domains.*

where $a_\nu(\cdot\,,\cdot)$ is the sesquilinear form

$$\begin{aligned}a_\nu(X,Y) = &\int_\Omega \nabla\varphi.\nabla\bar\psi\, d\Omega + \int_B \alpha\, d_{x_1}u_1\, d_{x_1}\bar v_1\, dB + \int_B \beta\, d^2_{x_1}u_2\, d^2_{x_1}\bar v_2\, dB\\ &+\sum_{k=1}^{2}(\mathbb{K}_k s_k).\bar t_k - \nu\left(\int_S \varphi\bar\psi\, dS + \int_B \rho\, u.\bar v\, dB + \sum_{k=1}^{2}(\mathrm{M}_k s_k).\bar t_k\right)\\ &-\nu^{1/2}\sum_{k=1}^{2}\int_{\Gamma_k}(s_k.n_g\,\bar\psi + \varphi\,\bar t_k.n_g)\, d\Gamma_k,\end{aligned}$$

and $l(f;\cdot)$ is the semilinear form

$$l(f;Y) = -\sum_{k=1}^{2}\left(\int_{\Gamma_k} f'_k\bar\psi\, d\Gamma_k + f''_k.\bar t_k\right),$$

where $Y = (\psi, v) \in \mathcal{H}$ and t_k is defined as in (5.1) by

$$t_k = (v_1(O_k), v_2(O_k), d_{x_1}v_2(O_k)).$$

Using the same techniques as in the proof of Proposition 2.1, it may be easily seen that a_ν is a continuous and coercive form on $\mathcal{H}\times\mathcal{H}$. The statement of Proposition 5.1 then follows from Lax–Milgram's theorem. □

5.2. Reduction to a bounded domain. Let $\hat\Omega \subset \Omega$ be the bounded domain delimited by two vertical segments Σ_1 and Σ_2 (located, respectively, at $x_1 = a_1$ and $x_1 = a_2$) chosen such that the hulls Γ_1 and Γ_2 are contained in the boundary $\partial\hat\Omega$ of $\hat\Omega$ (see Fig. 5.3). We denote by $\hat S$ and $\hat F$ the parts of S and F which are included in $\partial\hat\Omega$, and by $\check\Omega_j$, $\check S_j$, and $\check F_j$ (for $j = 1, 2$) the left and right connected components of $\Omega\setminus\hat\Omega$, $S\setminus\hat S$ and $F\setminus\hat F$, respectively.

As in § 2.5, we define, for $j = 1, 2$ and $\nu \in \mathbb{R}^+ \cup \mathbb{C}^+$, the coupling operator $\mathcal{Q}_{j,\nu}$ from $H^{1/2}(\Sigma_j)$ to $H^{-1/2}(\Sigma_j)$ by

$$\mathcal{Q}_{j,\nu}\,\tau = \partial_n(\check{\mathcal{R}}_{j,\nu}\,\tau)_{|\Sigma_j}, \tag{5.17}$$

where $\check{\mathcal{R}}_{j,\nu}\,\tau = \check\varphi$ is the solution of the following outer Dirichlet problem $\check{\mathcal{P}}_{j,\nu}$:

$$\left\{\begin{array}{l}\text{Find } \check\varphi \text{ in } H^1_{\text{loc}}(\check\Omega_j) \text{ if } \nu\in\mathbb{R}^+ \text{ or } H^1(\check\Omega_j) \text{ if } \nu\in\mathbb{C}^+ \text{ such that}\\ \Delta\check\varphi = 0 \quad\text{in } \check\Omega_j,\\ \partial_n\check\varphi - \nu\,\check\varphi = 0 \quad\text{on } \check S_j,\\ \partial_n\check\varphi = 0 \quad\text{on } \check F_j,\\ \check\varphi = \tau \quad\text{on } \Sigma_j,\\ \check\varphi \text{ satisfies the radiation condition (5.10)} \quad\text{if } \nu\in\mathbb{R}^+.\end{array}\right. \tag{5.18}$$

The well-posedness of this latter problem, as well as the continuity of the solution operator $\check{\mathcal{R}}_{j,\nu}$, may be proved by the same method as for Proposition 2.2. Consequently, $\mathcal{Q}_{j,\nu}$ is continuous from $H^{1/2}(\Sigma_j)$ to $H^{-1/2}(\Sigma_j)$.

Consider then the reduced problem $\hat{\mathcal{P}}_\nu$, set in the bounded fluid domain $\hat{\Omega}$:

$$\begin{array}{l}\text{Find } \hat{X} = (\hat{\varphi}, \hat{u}) \in \hat{\mathcal{H}} \text{ such that} \\ \Delta\hat{\varphi} = 0 \quad \text{in } \hat{\Omega}, \\ \partial_n\hat{\varphi} - \nu\,\hat{\varphi} = 0 \quad \text{on } \hat{S}, \\ \partial_n\hat{\varphi} = 0 \quad \text{on } \hat{F}, \\ \nu\rho\,\hat{u}_1 + d_{x_1}(\alpha\, d_{x_1}\hat{u}_1) = 0 \quad \text{on } B, \\ \nu\rho\,\hat{u}_2 - d^2_{x_1}(\beta\, d^2_{x_1}\hat{u}_2) = 0 \quad \text{on } B, \\ \nu^{1/2}\,\hat{s}_k.n_g - \partial_n\hat{\varphi} = f'_k \quad \text{on } \Gamma_k, \\ (\nu\,\mathbb{M}_k - \mathbb{K}_k)\hat{s}_k + \nu^{1/2}\int_{\Gamma_k}\hat{\varphi}\, n_g\, d\Gamma_k + \hat{g}_k = f''_k, \\ \partial_n\hat{\varphi} = -\mathcal{Q}_{j,\nu}\,\hat{\varphi}_{|\Sigma_j} \quad \text{on } \Sigma_j,\end{array} \tag{5.19}$$

where $\hat{s}_k$ and $\hat{g}_k$ are defined as in (5.1) and (5.9), and $\hat{\mathcal{H}}$ denotes the Hilbert space

$$\hat{\mathcal{H}} = H^1(\hat{\Omega}) \times (H^1(B) \times H^2(B)). \tag{5.20}$$

The equivalence between $\hat{\mathcal{P}}_\nu$ and the initial problem $\mathcal{P}_\nu$, which is expressed in Proposition 5.2 below, follows from the same arguments as for Proposition 2.3.

PROPOSITION 5.2. *Let $\nu \in \mathbb{R}^+ \cup \mathbb{C}^+$; for every $f \in \mathcal{F}$, problem $\mathcal{P}_\nu$ has at least (respectively, at most) one solution if and only if the same holds for $\hat{\mathcal{P}}_\nu$. Moreover, if $\hat{X} = (\hat{\varphi}, \hat{u})$ is a solution of $\hat{\mathcal{P}}_\nu$, then $X = (\varphi, \hat{u})$, where φ is the function given by*

$$\varphi_{|\hat{\Omega}} = \hat{\varphi} \quad \textit{and} \quad \varphi_{|\check{\Omega}} = \check{\mathcal{R}}_\nu\,\hat{\varphi}_{|\Sigma},$$

is a solution of $\mathcal{P}_\nu$. Conversely, if $X = (\varphi, u)$ is a solution of $\mathcal{P}_\nu$, then $\hat{X} = (\varphi_{|\hat{\Omega}}, u)$ is a solution of $\hat{\mathcal{P}}_\nu$.

5.3. Analytic continuation. The analytic extension of the reduced problem $\hat{\mathcal{P}}_\nu$ is based on the explicit series expansion of the coupling operators $\mathcal{Q}_{j,\nu}$ which is achieved, here again, by the method of separation of variables.

In the case of real positive ν, it is an easy matter to check that the solutions with separated variables of the outer problem $\check{\mathcal{P}}_{j,\nu}$ are given (up to a multiplicative complex constant) by

$$\check{\varphi}^{(m)}_{j,\nu}(x_1, x_2) = \tau^{(m)}_\nu(x_2)\; e^{(-1)^{j+1}\zeta^{(m)}_\nu(x_1 - a_j)} \quad \text{for } m \geq 0, \tag{5.21}$$

where the functions $\tau^{(m)}_\nu$ and the complex numbers $\zeta^{(m)}_\nu$ are defined as in Lemma 3.1. Note that $\check{\varphi}^{(0)}_{j,\nu}$ is the only radiative solution: the others are evanescent, since they decrease exponentially at infinity.

Using some classical results of the spectral theory of selfadjoint operators (see the proof of Proposition 3.1), we readily prove the following.

PROPOSITION 5.3. *Let $\nu > 0$; the set $\mathcal{T}_\nu = \{\tau^{(m)}_\nu;\ m \geq 0\}$ is an orthonormal basis of $L^2(\Sigma_j)$ and an orthogonal basis of $H^s(\Sigma_j)$ for every $s \in \,]0, 1]$. Furthermore, the expression*

$$[\tau]_{\nu,s} = \left(\sum_{m \geq 0} (1 + m^2)^s \left| (\tau, \tau^{(m)}_\nu)_{\Sigma_j} \right|^2 \right)^{1/2} \tag{5.22}$$

is a norm on $H^s(\Sigma_j)$ equivalent to the usual norm.

As a consequence, every function $\tau \in H^{1/2}(\Sigma_j)$ expands as

$$\tau = \sum_{m\geq 0} (\tau, \tau_\nu^{(m)})_{\Sigma_j}\, \tau_\nu^{(m)}.$$

Noticing that $\check{\varphi}_{j,\nu}^{(m)}$ (given by (5.21)) is nothing but the solution of $\check{\mathcal{P}}_{j,\nu}$ for the Dirichlet datum $\tau_\nu^{(m)}$, we infer that the general solution of $\check{\mathcal{P}}_{j,\nu}$ reads

$$\check{\mathcal{R}}_{j,\nu}\,\tau = \sum_{m\geq 0} (\tau, \tau_\nu^{(m)})_{\Sigma_j}\, \check{\varphi}_{j,\nu}^{(m)},$$

where the series converges in $H^1_{\text{loc}}(\check{\Omega}_j)$. The diagonal form of the coupling operator for $\nu \in \mathbb{R}^+$ follows:

$$Q_{j,\nu}\,\tau = \sum_{m\geq 0} \zeta_\nu^{(m)}\, (\tau, \tau_\nu^{(m)})_{\Sigma_j}\, \tau_\nu^{(m)} \quad \forall \tau \in H^{1/2}(\Sigma_j). \tag{5.23}$$

The case of complex ν is dealt with as described in §3.2. We have seen that each solution $\zeta_\nu^{(m)}$ of the dispersion equation extends analytically in the simply connected domain $\mathbb{D} \subset \mathbb{C}$. The same holds for the functions $\tau_\nu^{(m)}$ and $\check{\varphi}_{j,\nu}^{(m)}$ (which are defined everywhere in $\mathbb{D}$). Note that if $\nu \in \mathbb{D} \setminus \mathbb{R}^+$, each solution $\check{\varphi}_{j,\nu}^{(m)}$ decreases exponentially at infinity except in the case $m = 0$ and $\nu \in \mathbb{D}^-$, where it becomes exponentially increasing (see Proposition B.5). Proposition 5.3 extends as follows.

PROPOSITION 5.4. *For every $\nu \in \mathbb{D}$ and every $s \in [0,1]$, the set $\mathcal{T}_\nu = \{\tau_\nu^{(m)};\ m \geq 0\}$ is a basis of $H^s(\Sigma_j)$ and the expression $[\tau]_{\nu,s}$ given by (5.22) is still a norm on $H^s(\Sigma_j)$ equivalent to the usual norm. Moreover, the two families $\mathcal{T}_\nu$ and $\mathcal{T}_{\bar{\nu}}$ are adjoint to each other in $L^2(\Sigma_j)$.*

The proof of this statement is simpler than Theorem 3.1. We use directly the perturbation result given in the proof of Lemma 3.3. Indeed, it may be easily seen that for every $\nu \in \mathbb{D}$ and $\mu \in \mathbb{R}^+$, we have

$$\sum_{m\geq 0} \frac{\left[\tau_\nu^{(m)} - \tau_\mu^{(m)}\right]^2_{\mu,s}}{\left[\tau_\mu^{(m)}\right]^2_{\mu,s}} < \infty.$$

We readily deduce from Proposition 5.4 that if $\nu \in \mathbb{D}^+$, the coupling operator $Q_{j,\nu}$ expands as

$$Q_{j,\nu}\,\tau = \sum_{m\geq 0} \zeta_\nu^{(m)}\, (\tau, \tau_{\bar{\nu}}^{(m)})_{\Sigma_j}\, \tau_\nu^{(m)} \quad \forall \tau \in H^{1/2}(\Sigma_j). \tag{5.24}$$

This expansion actually defines the analytic continuation of $Q_{j,\nu}$: it appears as an analytic family from $\mathbb{D}$ to $\mathcal{B}(H^{1/2}(\Sigma_j), H^{-1/2}(\Sigma_j))$.

We can then proceed to the analytic extension of the reduced problem $\hat{\mathcal{P}}_\nu$ by writing it in variational form:

$$\begin{aligned} &\text{Find } \hat{X} = (\hat{\varphi}, \hat{u}) \in \hat{\mathcal{H}} \text{ such that} \\ &\hat{a}_\nu(\hat{X}, \hat{Y}) = \hat{l}(f; \hat{Y}) \quad \forall \hat{Y} \in \hat{\mathcal{H}}, \end{aligned} \tag{5.25}$$

where $\hat{a}_\nu$ is the sesquilinear form defined on $\hat{\mathcal{H}} \times \hat{\mathcal{H}}$ by

$$\hat{a}_\nu(\hat{X},\hat{Y}) = \hat{k}(\hat{X},\hat{Y}) - \nu\, \hat{m}(\hat{X},\hat{Y}) - \nu^{1/2}\, \hat{c}(\hat{X},\hat{Y}) + \hat{q}_\nu(\hat{X},\hat{Y}), \tag{5.26}$$

with

$$\hat{k}(\hat{X},\hat{Y}) = \int_{\hat{\Omega}} \nabla\hat{\varphi}.\overline{\nabla\hat{\psi}}\, d\hat{\Omega} + \int_B \alpha\, d_{x_1}\hat{u}_1\overline{d_{x_1}\hat{v}_1}\, dB + \int_B \beta\, d^2_{x_1}\hat{u}_2\overline{d^2_{x_1}\hat{v}_2}\, dB + \sum_{k=1}^{2}(\mathbb{K}_k\hat{s}_k).\overline{\hat{t}_k}, \tag{5.27}$$

$$\hat{m}(\hat{X},\hat{Y}) = \int_{\hat{S}} \hat{\varphi}\overline{\hat{\psi}}\, d\hat{S} + \int_B \rho\, \hat{u}.\overline{\hat{v}}\, dB + \sum_{k=1}^{2}(\mathbb{M}_k\hat{s}_k).\overline{\hat{t}_k}, \tag{5.28}$$

$$\hat{c}(\hat{X},\hat{Y}) = \sum_{k=1}^{2}\int_{\Gamma_k} (\hat{s}_k.n_g\,\overline{\hat{\psi}} + \hat{\varphi}\,\overline{\hat{t}_k.n_g})\, d\Gamma_k, \tag{5.29}$$

$$\hat{q}_\nu(\hat{X},\hat{Y}) = \sum_{j=1}^{2}\sum_{m\geq 0} \zeta_\nu^{(m)}\, (\hat{\varphi}, \tau_{\bar{\nu}}^{(m)})_{\Sigma_j}\, (\tau_\nu^{(m)}, \hat{\psi})_{\Sigma_j} \tag{5.30}$$

for all $\hat{X} = (\hat{\varphi}, \hat{u})$ and $\hat{Y} = (\hat{\psi}, \hat{v})$ in $\hat{\mathcal{H}}$. The semilinear form $\hat{l}$ is given by

$$\hat{l}(f;\hat{Y}) = -\sum_{k=1}^{2}\left(\int_{\Gamma_k} f'_k\overline{\hat{\psi}}\, d\Gamma_k + f''_k.\overline{\hat{t}_k}\right). \tag{5.31}$$

Let $\hat{A}_\nu$ denote the bounded operator on $\hat{\mathcal{H}}$ associated with $\hat{a}_\nu$, i.e., defined by the relation $(\hat{A}_\nu\hat{X},\hat{Y})_{\hat{\mathcal{H}}} = \hat{a}_\nu(\hat{X},\hat{Y})$. As in Theorem 4.1, a characterization of the eigenvalues and scattering frequencies of the problem is as follows.

THEOREM 5.1. *The operators $\hat{A}_\nu$ for $\nu \in \mathbb{D}$ form a holomorphic family of Fredholm operators on $\hat{\mathcal{H}}$. Moreover, $\hat{A}_\nu^{-1}$ is a meromorphic family in $\mathbb{D}$ whose poles are located in $\mathbb{R}^+$ or $\mathbb{D}^-$. These poles are the solutions of the following nonlinear* (*and nonselfadjoint*) *eigenvalue problem:*

$$\begin{aligned}&\textit{Find } \nu \in \mathbb{R}^+ \cup \mathbb{D}^- \textit{ such that there exists } \hat{X} \in \hat{\mathcal{H}} \setminus \{0\}, \textit{ which satisfies}\\ &\hat{k}(\hat{X},\hat{Y}) - \nu^{1/2}\, \hat{c}(\hat{X},\hat{Y}) + \hat{q}_\nu(\hat{X},\hat{Y}) = \nu\, \hat{m}(\hat{X},\hat{Y}) \quad \forall \hat{Y} \in \hat{\mathcal{H}}.\end{aligned} \tag{5.32}$$

5.4. Numerical implementation. The discretization of the problem involves two steps. We first truncate the series expansion of the sesquilinear form $\hat{q}_\nu$ (we denote by M the order of truncation of the series). We then define a finite-dimensional subspace of $\hat{\mathcal{H}}$ by means of a standard conforming element method. For the fluid domain $\hat{\Omega}$, we use classical quadrangular Lagrange finite elements (see, e.g., Ciarlet [3]), i.e., a piecewise polynomial interpolation of partial order 1 (class $\mathcal{C}^0$) for $\hat{\varphi}$. For the beam, we use standard beam elements, i.e., a polynomial interpolation of order 1 (class $\mathcal{C}^0$) for $\hat{u}_1$, and of order 3 (class $\mathcal{C}^1$) for $\hat{u}_2$. Let $\mathcal{H}_h$ denote the approximation space resulting from this finite element discretization, and let $(\cdot \mid \cdot)_h$ be the scalar product in $\mathcal{H}_h$. Consider then the matrices $\mathbb{M}_h$, $\mathbb{K}_h$, $\mathbb{C}_h$, and $\mathbb{Q}^{(M)}_{\nu,h}$ given by

$$\begin{aligned}(\mathbb{M}_h X_h \mid Y_h)_h &= \hat{m}(X_h, Y_h),\\ (\mathbb{K}_h X_h \mid Y_h)_h &= \hat{k}(X_h, Y_h),\\ (\mathbb{C}_h X_h \mid Y_h)_h &= \hat{c}(X_h, Y_h),\\ (\mathbb{Q}^{(M)}_{\nu,h} X_h \mid Y_h)_h &= \sum_{j=1}^{2}\sum_{m=0}^{M} \zeta_\nu^{(m)}\, (\varphi_h, \tau_{\bar{\nu}}^{(m)})_{\Sigma_j}\, (\tau_\nu^{(m)}, \psi_h)_{\Sigma_j}\end{aligned}$$

for all $X_h = (\varphi_h, u_h)$ and $Y_h = (\psi_h, v_h)$ in $\mathcal{H}_h$. Note that $\mathbb{M}_h$, $\mathbb{K}_h$, and $\mathbb{C}_h$ are real symmetric matrices, and $\mathbb{Q}_{\nu,h}^{(M)}$ is complex symmetric (but not selfadjoint). These matrices are computed by classical assembly techniques. For the sake of simplicity, we omit the index "h" in the sequel.

The numerical application concerns three points: the usual approach which consists in computing the response curve of the body for real frequencies, the present method of determination of scattering frequencies, and finally, a comparison between both approaches, by means of the series expansion of the response in the vicinity of the scattering frequencies.

(i) The computation of the response curve of the floating body is performed as follows. Consider a monochromatic incident wave of frequency $\nu^{1/2} \in \mathbb{R}^+$ given by

$$\varphi_I(x_1, x_2) = \frac{\cosh \nu_0(x_2+1)}{\nu^{1/2} \cosh \nu_0}\, e^{i\nu_0 x_1},$$

where ν_0 is the positive root of (5.11) (the amplitude of this wave is 1, since the free surface elevation is defined by $\eta(x_1) = i\,\nu^{1/2}\varphi_I(x_1, 0)$; see [10]). From (5.25), the approximate response of the system is then the solution $X = (\varphi, u)$ of the following linear problem:

$$(\mathbb{K} - \nu\mathbb{M} - \nu^{1/2}\mathbb{C} + \mathbb{Q}_\nu^{(M)})\, X = F_\nu, \tag{5.33}$$

where the right-hand member is defined by

$$(F_\nu \mid Y) = -\sum_{k=1}^{2} \int_{\Gamma_k} (\partial_n \varphi_I\, \bar{\psi} - \nu^{1/2}\varphi_I\, \bar{t}_k.n_g)\, d\Gamma_k. \tag{5.34}$$

The total energy of the catamaran follows:

$$E_\nu = \tfrac{1}{2}((\nu\mathbb{M} + \mathbb{K})\dot{X} \mid \dot{X}), \quad \text{where } \dot{X} = (0, u). \tag{5.35}$$

By computing this quantity step by step over a given frequency range, we obtain the response curve of the catamaran: its peaks are the resonant states of the system.

(ii) On the other hand, consider now the matrix nonlinear eigenvalue problem associated with (5.32),

$$(\mathbb{K} - \nu^{1/2}\mathbb{C} + \mathbb{Q}_\nu^{(M)})\, X = \nu\mathbb{M}\, X, \tag{5.36}$$

whose solutions are the approximate scattering frequencies. Solving this problem amounts to determining the solutions of the fixed-point equation $\lambda_k(\nu) = \nu$, where $\lambda_k(\nu)$ denotes any eigenvalue of the problem

$$(\mathbb{K} - \nu^{1/2}\mathbb{C} + \mathbb{Q}_\nu^{(M)})\, X = \lambda_k(\nu)\, \mathbb{M}\, X. \tag{5.37}$$

This latter equation is solved as described in [18, §6.3] by an iterative Newton method.

(iii) As for the continuous problem, the solution of (5.33) is a meromorphic function of ν. In the vicinity of an approximate scattering frequency ν_*, it expands as

$$X = \sum_{p \geq -P} (\nu - \nu_*)^p X^{(p)}, \qquad P \geq 0.$$

We described in [18] how to calculate the first terms of this expansion. In particular, if ν_* is a nonsingular simple scattering frequency (i.e., $\lambda_k(\nu_*) = \nu_*$ is a simple eigenvalue of (5.37) and $d_\nu \lambda_k(\nu_*) \neq 1$), we have $P = 1$ and

$$X^{(-1)} = (\varphi^{(-1)}, u^{(-1)}) = \frac{(F^{(0)} \mid \bar{X}_{\nu_*})}{((\frac{1}{2}\,\nu_*^{-1/2}\mathbb{C} + d_\nu \mathbb{Q}_{\nu_*}^{(M)} - \mathbb{M})X_{\nu_*} \mid \bar{X}_{\nu_*})}\, X_{\nu_*}, \tag{5.38}$$

where $F^{(0)}$ is the first term of the expansion of F_ν near ν_* (F_ν depends analytically on ν), and X_{ν_*} is a scattering mode associated with ν_* (i.e., an eigenvector of (5.37) for $\nu = \nu_*$ associated with the eigenvalue $\lambda_k(\nu_*) = \nu_*$). In this case, the total energy of the catamaran expands as

$$E_\nu = \frac{|\nu - \nu_*|^{-2}}{2}\, ((\nu_* \mathbb{M} + \mathbb{K})\dot{X}^{(-1)} \mid \dot{X}^{(-1)}) + O(|\nu - \nu_*|^{-1}), \tag{5.39}$$

where $\dot{X}^{(-1)} = (0, u^{(-1)})$. For the scattering frequencies which are close enough to the positive real axis, we will see that the only first term of this expansion provides a good approximation of the response curve of the catamaran.

5.5. Numerical results. We present a typical numerical application which is related to the following data. The hulls of the two rigid bodies are parabolic curves defined by

$$x_2 = 2(x_1 \pm 1)^2 - 0.2\,.$$

The centres of gravity of the bodies are $G_k = (\pm 1, 0)$, and the embedding points are $O_k = (\pm 1, 0.3)$. The inertia of each body at point O_k is 0.2. The mechanical data of the beam are assumed constant: $\rho = 0.04$, $\alpha = 0.7$, and $\beta = 0.35$.

The discretizetion of the fluid domain is shown in Fig. 5.4 (770 degrees of freedom for φ). For the beam, we use 38 elements (117 degrees of freedom for u).

The upper part of Fig. 5.5 shows the response curve of the catamaran (i.e., function E_ν defined in (5.35)) in the frequency range $[0.7, 3.0]$. In the lower part of the same figure, the locations of the computed scattering frequencies in the lower complex half plane (solutions of (5.36)) are displayed. All these scattering frequencies are simple; their numerical values are

	(a)	(b)	(c)	(d)	(e)	(f)	(g)
real part	1.028	1.306	1.636	1.776	2.186	2.530	2.723
imag. part	0.175	0.015	0.191	0.011	0.143	0.043	0.010.

As expected, their real parts agree with the resonant states of the catamaran, i.e. the peaks of the response curve. Note that the sharp peaks correspond to the scattering frequencies which are very close to the positive real axis (imaginary part about 10^{-2} for (b), (d), and (g)).

Figs. 5.6(a)–5.6(e) show the scattering modes associated with the scattering frequencies (a)–(e). Only the strain of the catamaran and the free surface elevation are represented. The location of the system at rest is indicated by the dotted line.

Case (a) is nearly a rigid motion of the catamaran (very slight strain of the beam): it is essentially a "rool" motion coupled with a small "sway" (displacement in the x_1 direction).

Case (b) shows, on the other hand, an important bending strain of the beam with nearly no global motion of the catamaran. In fact, this mode is almost an eigenmode

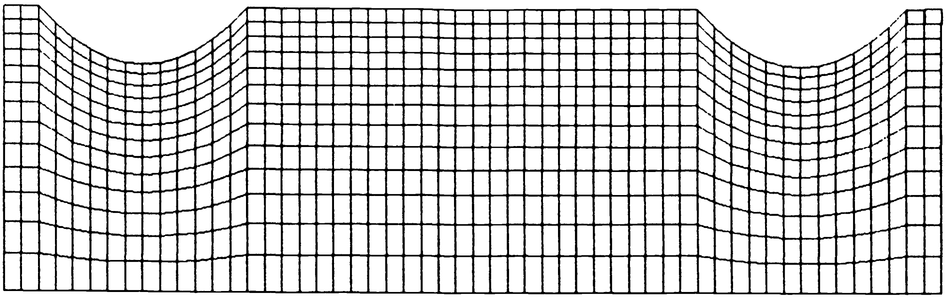


FIG. 5.4. *Discretization of the fluid domain.*

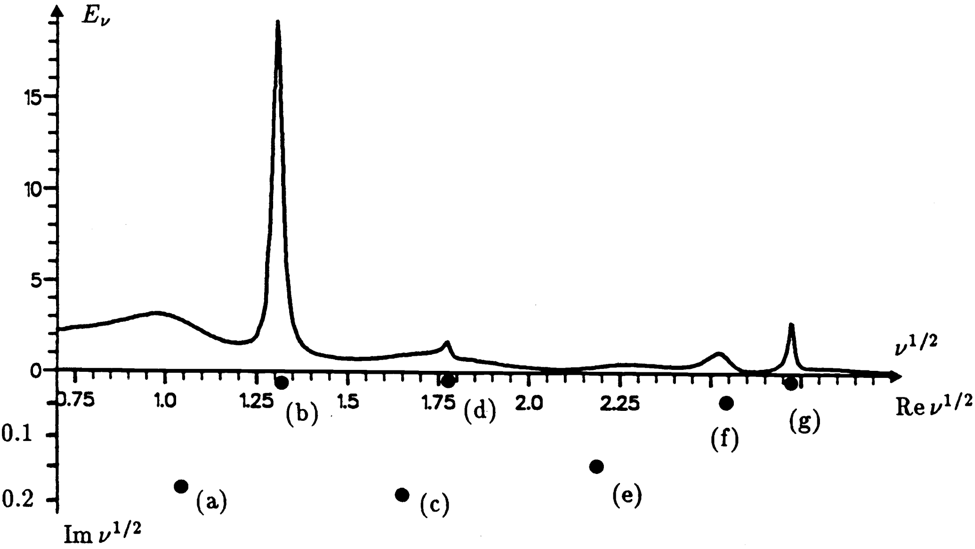


FIG. 5.5. *Response curve and scattering frequencies.*

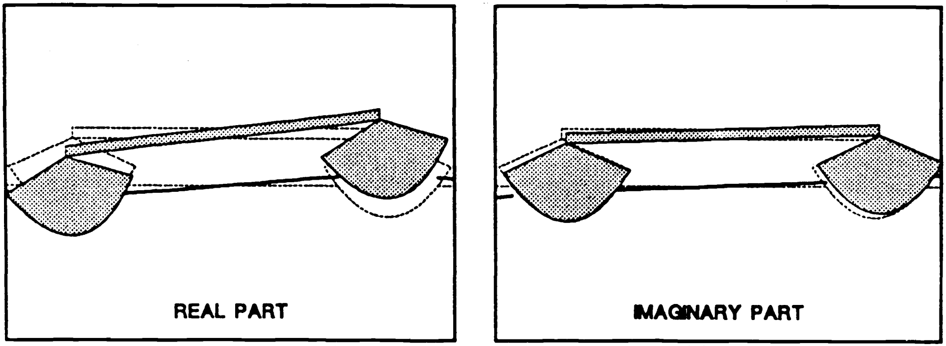


FIG. 5.6(a). *Scattering mode associated with* (a).

of the body uncoupled with the fluid, since the free surface elevation is small with respect to the displacement of the beam.

Case (c) is a symmetric "heave" motion (vertical global displacement) together with a slight longitudinal and bending strain of the beam. As in case (a), this is clearly a coupled mode between the fluid and the floating body.

Case (d) shows a large free surface elevation compared with the motion of the catamaran. In fact, this elevation is nearly zero outside the two hulls: the wave is "trapped" between them. Contrary to (b), this mode is almost a scattering mode (or a "trapped mode") of the fluid uncoupled with the floating body. This explains in particular why the associated peak of the response curve is low.

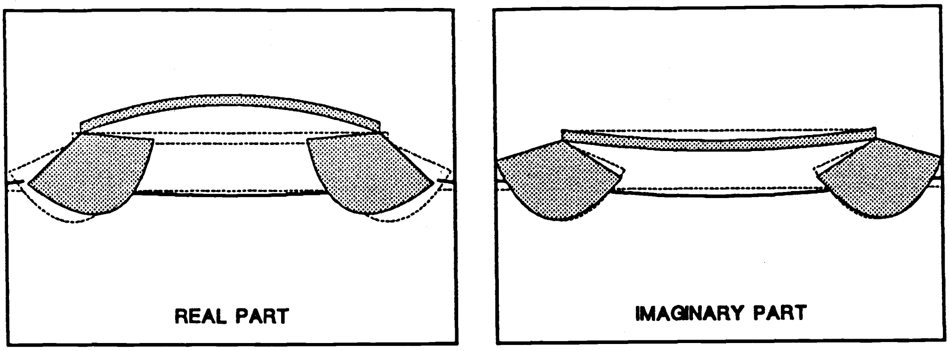


FIG. 5.6(b). *Scattering mode associated with* (b).

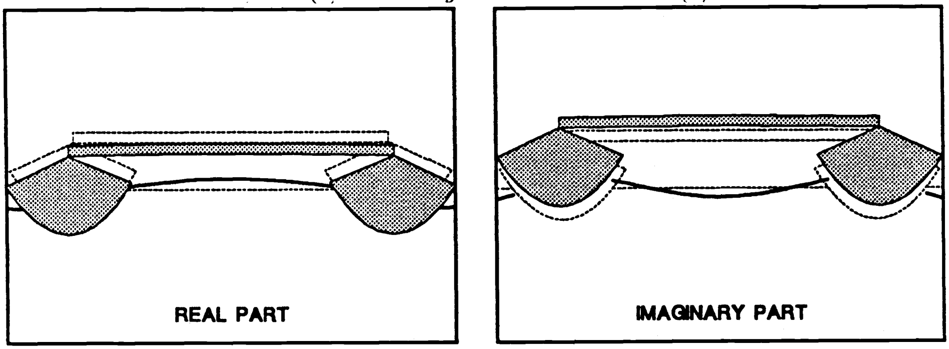


FIG. 5.6(c). *Scattering mode associated with* (c).

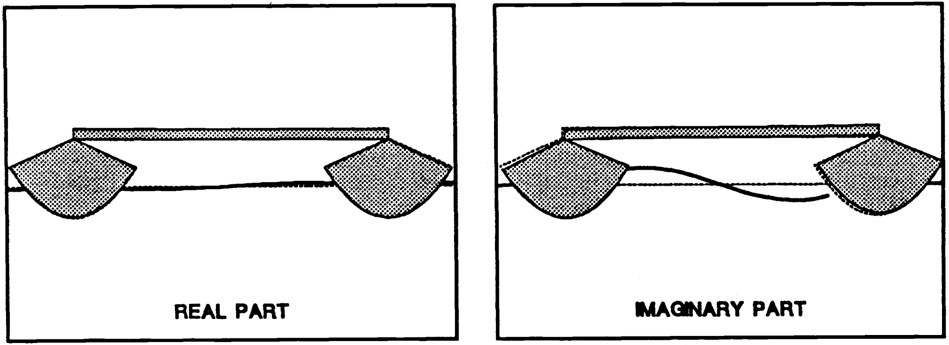


FIG. 5.6(d). *Scattering mode associated with* (d).

Case (e) is a coupled scattering mode of the system: as in case (c), it essentially consists in a global heave motion and a longitudinal strain of the beam.

Fig. 5.7 compares the response curve of the catamaran (i.e., the curve of Fig 5.5, reproduced here in dotted line, with the approximation of resonant states which follows from the first term of the expansion of the response (order -1) in the vicinity of the scattering frequency (i.e., formulas (5.38) and (5.39) applied for real frequencies). This approximation is rather good for (b) and (g) (relative error about 10^{-2} for the maximum). In case (d), the location of the peak is correct but not its height: intuitively, this may be accounted for by the fact that the energy of the catamaran in the scattering mode (d) is low. For the other resonant states, we notice a sensible shift between the peak and its approximation: the only first term of the expansion of the response is not sufficient to provide a precise location of the peak (an estimate of this shift using the second term of the expansion is given in [18]).

Appendix A. Series expansion of the solution of the outer problem. This

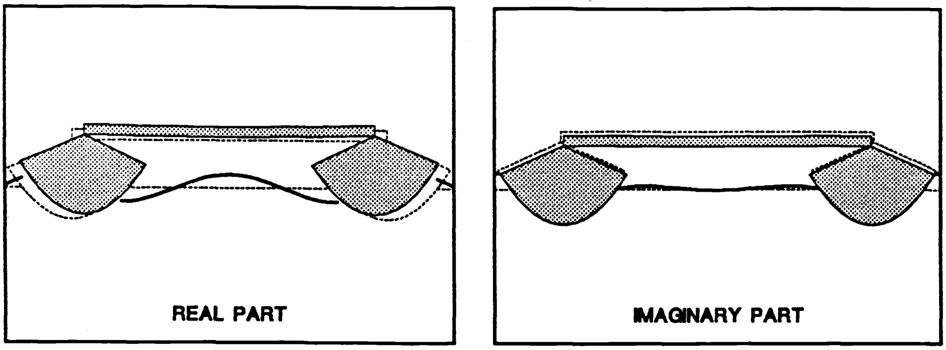


FIG. 5.6(e). *Scattering mode associated with* (e).

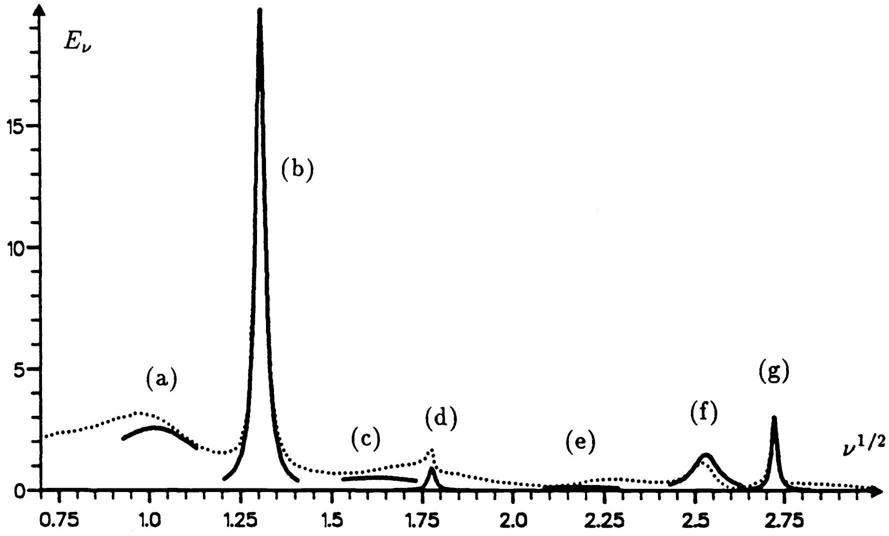


FIG. 5.7. *Approximation of resonant states.*

appendix is devoted to the proof of Lemma 3.2. We show that the series constructed in §3.1 for $\nu \in \mathbb{R}^+$ (see (3.20))

$$\check{\varphi} = \sum_{n\in\mathbb{Z}} \sum_{m\geq 0} (\chi, \chi_\nu^{(n,m)})_\Sigma \, \check{\varphi}_\nu^{(n,m)} \tag{A.1}$$

converges in $H^1_{\text{loc}}(\check{\Omega})$ for every $\chi \in H^{1/2}(\Sigma)$, depends continuously on χ and satisfies the radiation condition (2.14). These results are based on asymptotic properties of the modified Bessel functions K_n which are collected in §A.3.

Let us point out that the method we present below can be used similarly for the two-dimensional acoustic scattering problem, in order to prove that the Fourier series expansion of the scattered wave outside a circle satisfies the Sommerfeld radiation condition. This answers the question raised by Hochstadt [8].

A.1. Convergence of the series. Consider, for a given $R > r_0$, the bounded domain $\check{\Omega}_R \subset \check{\Omega}$ delimited by the vertical cylinder $\Sigma_R = \{x \in \check{\Omega};\ x_1^2 + x_2^2 = R^2\}$; we denote by $\check{S}_R$ the part of $\check{S}$ which is contained in the boundary of $\check{\Omega}_R$. The proof of the convergence of the series (A.1) in $H^1(\check{\Omega}_R)$ is based on a suitable choice of a scalar product for which the solutions with separated variables become orthogonal.

As in the proof of Proposition 3.1, it may be easily seen that the expression

$$N_{d,R}(\check\varphi) = \left(\|\nabla\check\varphi\|^2_{\check\Omega_R} + d\,\|\check\varphi\|^2_{\check\Omega_R} - \nu\,\|\check\varphi\|^2_{\check S_R}\right)^{1/2} \tag{A.2}$$

is, for sufficiently large $d > 0$, a norm on $H^1(\check\Omega_R)$ equivalent to the usual norm. Let $P_{d,R}(\cdot\,,\cdot)$ denote the associated scalar product. Noticing that if $\check\varphi$ is a solution of $\check{\mathcal{P}}_\nu$, we have

$$(\nabla\check\varphi,\nabla\psi)_{\check\Omega_R} - \nu\,(\check\varphi,\psi)_{\check S_R} = \langle\partial_n\check\varphi,\psi\rangle_{1/2,\Sigma\cup\Sigma_R} \quad \forall\psi\in H^1(\check\Omega_R),$$

where $\langle\cdot\,,\cdot\rangle_{1/2,\Sigma\cup\Sigma_R}$ denotes the duality product between $H^{-1/2}(\Sigma\cup\Sigma_R)$ and $H^{1/2}(\Sigma\cup\Sigma_R)$, we deduce from the definition (3.1) of $\check\varphi_\nu^{(n,m)}$ that

(A.3)
$$\begin{aligned}P_{d,R}(\check\varphi_\nu^{(n,m)},\check\varphi_\nu^{(n',m')}) =&(\chi_\nu^{(n,m)},\chi_\nu^{(n',m')})_\Sigma\\ &\times\left(d\int_{r_0}^{R}\eta_\nu^{(n,m)}\overline{\eta_\nu^{(n',m')}}\,\frac{r}{r_0}\,dr + \left[\frac{r}{r_0}\,d_r\eta_\nu^{(n,m)}\overline{\eta_\nu^{(n',m')}}\right]_{r_0}^{R}\right),\end{aligned}$$

which shows that functions $\check\varphi_\nu^{(n,m)}$ form an orthogonal family in $H^1(\check\Omega_R)$ for $P_{d,R}$. Consequently, the norm of (A.1) reads

$$N_{d,R}(\check\varphi) = \left(\sum_{n\in\mathbb{Z}}\sum_{m\geq 0}\left|(\chi,\chi_\nu^{(n,m)})_\Sigma\right|^2 N^2_{d,R}(\check\varphi_\nu^{(n,m)})\right)^{1/2}. \tag{A.4}$$

By Proposition 3.1, if there exists $C > 0$, independent of n and m, such that

$$N^2_{d,R}(\check\varphi_\nu^{(n,m)}) \leq C\,(1+n^2+m^2)^{1/2}, \tag{A.5}$$

then $N_{d,R}(\check\varphi)$ is bounded and satisfies the following inequality:

$$N_{d,R}(\check\varphi) \leq \sqrt{C}\ [\chi]_{\nu,1/2}, \tag{A.6}$$

which is nothing but the required continuity property (2.26). From (A.3), we thus have to estimate the asymptotic behaviour, when $n^2+m^2\to+\infty$, of

$$N^2_{d,R}(\check\varphi_\nu^{(n,m)}) = d\int_{r_0}^{R}\left|\eta_\nu^{(n,m)}\right|^2\frac{r}{r_0}\,dr + \left[\frac{r}{r_0}\,d_r\eta_\nu^{(n,m)}\overline{\eta_\nu^{(n,m)}}\right]_{r_0}^{R}, \tag{A.7}$$

where

$$\eta_\nu^{(n,m)}(r) = \frac{K_n(\zeta_\nu^{(m)}r)}{K_n(\zeta_\nu^{(m)}r_0)} \quad\text{and}\quad d_r\eta_\nu^{(n,m)}(r) = \frac{\zeta_\nu^{(m)}K_n'(\zeta_\nu^{(m)}r)}{K_n(\zeta_\nu^{(m)}r_0)}. \tag{A.8}$$

Note that, from the symmetry property (3.10), it is enough to consider the case $n\geq 0$.

First, assume that $m > 0$. In this case, the roots $\zeta_\nu^{(m)}$ of the dispersion equation are real and positive, and their asymptotic behaviour is given by (see §3.1):

$$\zeta_\nu^{(m)} \sim m\pi \quad\text{when } m\to+\infty. \tag{A.9}$$

The arguments of K_n and K_n' in (A.8) are thus real and positive. Setting $x = \zeta_\nu^{(m)} r$ and $x_0 = \zeta_\nu^{(m)} r_0$ in (A.24), we derive

(A.10)
$$\left|\frac{K_n'(\zeta_\nu^{(m)} r)}{K_n(\zeta_\nu^{(m)} r)}\right| \leq C\,\frac{(n^2+m^2)^{1/2}}{m} \quad \text{and} \quad \left|\frac{K_n(\zeta_\nu^{(m)} r)}{K_n(\zeta_\nu^{(m)} r_0)}\right| \leq C \quad \text{for } n \geq 0,\ m > 0.$$

These inequalities show that the first term of the right-hand side of (A.7) is bounded by a constant, and the second one, by $C(n^2+m^2)^{1/2}$. Relation (A.5) is thus proved.

Suppose now that $m = 0$. In this case, the arguments of K_n and K_n' in (A.8) are imaginary (for $\zeta_\nu^{(0)} = -i\nu_0$; see §3.1). Setting $y = \nu_0 r$ in (A.29) and (A.30), we see that

$$\text{(A.11)} \qquad \left|\frac{K_n'(-i\nu_0 r)}{K_n(-i\nu_0 r)}\right| \leq Cn \quad \text{and} \quad \left|\frac{K_n(-i\nu_0 r)}{K_n(-i\nu_0 r_0)}\right| \leq C \quad \text{for } r \in \,[r_0, R]\,,\quad n > 0.$$

These estimates show that inequality (A.5) is still valid for $m = 0$. The convergence of the series (A.1) follows.

A.2. The radiation condition. By construction, each term of the series (A.1) satisfies the radiation condition (2.14): the aim of this paragraph is to show that this property holds for the sum φ of the series. First, note that for $R > r_0$,

(A.12)
$$\int_{\Sigma_R} |\partial_r\varphi - i\nu_0\varphi|^2\, d\Sigma = \frac{R}{r_0}\sum_{n\in\mathbb{Z}}\sum_{m\geq 0}\left|(\chi, \chi_\nu^{(n,m)})_\Sigma\right|^2 \left|d_r\eta_\nu^{(n,m)}(R) - i\nu_0\eta_\nu^{(n,m)}(R)\right|^2,$$

since $\{\chi_\nu^{(n,m)};\ n\in\mathbb{Z},\ m\geq 0\}$ forms an orthonormal basis of $L^2(\Sigma)$. We prove below that

$$\text{(A.13)} \qquad \lim_{R\to+\infty} R\sum_{n\in\mathbb{Z}}\sum_{m\geq 0}\left|d_r\eta_\nu^{(n,m)}(R) - i\nu_0\eta_\nu^{(n,m)}(R)\right|^2 = 0,$$

which of course implies the convergence to 0 of (A.12) when $R \to +\infty$ for every $\chi \in H^{1/2}(\Sigma)$. Note that, here again, it is enough to deal with the case $n \geq 0$.

(i) In a first step, we consider the part of the series in (A.13) which corresponds to $m > 0$: the exponential decay of $\eta_\nu^{(n,m)}(r)$ actually leads to a stronger property than (A.13), namely,

$$\text{(A.14)} \qquad \lim_{R\to+\infty} R^\beta\sum_{n\geq 0}\sum_{m>0}\left|\eta_\nu^{(n,m)}(R)\right|^2 = 0 \quad \forall\beta > 0,$$

as well as the same property for $d_r\eta_\nu^{(n,m)}(R)$. We restrict ourselves to the proof of the former statement; the latter is obtained similarly. In order to prove (A.14), we study separately the two quantities:

$$\text{(A.15)} \qquad A_1(R) = R^\beta\sum_{m>0}\sum_{n=0}^{n_m(R)}\left|\eta_\nu^{(n,m)}(R)\right|^2,$$

$$\text{(A.16)} \qquad A_2(R) = R^\beta\sum_{m>0}\sum_{n>n_m(R)}\left|\eta_\nu^{(n,m)}(R)\right|^2,$$

where $n_m(R)$ denotes the greatest integer such that $n_m(R) \leq \zeta_\nu^{(m)} R$. If $n \leq n_m(R)$, we deduce from (A.25) that

$$\left|\eta_\nu^{(n,m)}(R)\right| = \left|\frac{K_n(\zeta_\nu^{(m)} R)}{K_n(\zeta_\nu^{(m)} r_0)}\right| \leq C\, e^{-\zeta_\nu^{(m)}(R-r_0)/3},$$

whence we infer, from (A.9), that for every $\gamma > 0$, there exists $C_\gamma > 0$ such that

$$\left|\eta_\nu^{(n,m)}(R)\right| \leq C_\gamma\ (mR)^{-\gamma} \quad \text{for } \ n \leq n_m(R).$$

Noticing that $n_m(R) \leq C\, mR$, we thus have

$$A_1(R) \leq C'_\gamma\ R^\beta \sum_{m>0} (mR)^{-2\gamma+1}.$$

By choosing, for instance, $\gamma = \beta + 1$, we deduce that $A_1(R) \to 0$ when $R \to +\infty$. On the other hand, if $n > n_m(R)$, inequality (A.26) shows that

$$\left|\eta_\nu^{(n,m)}(R)\right| \leq C \left(\frac{R}{2r_0}\right)^{-n}.$$

Noticing that $n_m(R)$ is an increasing function of m, we can rewrite the series (A.16) in the form

$$A_2(R) = R^\beta \sum_{n>n_1(R)} \ \sum_{m<m_n(R)} \left|\eta_\nu^{(n,m)}(R)\right|^2,$$

where $m_n(R)$ is the greatest value of m such that $\zeta_\nu^{(m)} R \leq n$. By (A.9), we see that $m_n(R) \leq C\, n/R$; consequently

$$A_2(R) \leq C'\, R^\beta \sum_{n>n_1(R)} \frac{n}{R} \left(\frac{R}{2r_0}\right)^{-2n}.$$

Hence, we easily deduce that $A_2(R) \to 0$, which completes the proof of (A.14).

(ii) Consider now the case $m = 0$ in the series (A.13): it remains to prove that

$$\text{(A.17)} \qquad \lim_{R\to+\infty} R \sum_{n>0} \left| d_r \eta_\nu^{(n,0)}(R) - i\nu_0 \eta_\nu^{(n,0)}(R)\right|^2 = 0,$$

where

$$\left| d_r \eta_\nu^{(n,0)}(R) - i\nu_0 \eta_\nu^{(n,0)}(R)\right| = \nu_0 \left| \frac{K'_n(-i\nu_0 R) + K_n(-i\nu_0 R)}{K_n(-i\nu_0 r_0)}\right|.$$

Let ε be an arbitrary small (but fixed) positive number. We split the series in (A.17) into three parts:

(A.18)

$$B_1(R) = R \sum_{n<n^-(R)} (...), \quad B_2(R) = R \sum_{n=n^-(R)}^{n^+(R)} (...), \quad B_3(R) = R \sum_{n>n^+(R)} (...),$$

where $n^-(R)$ and $n^+(R)$ denote, respectively, the greatest integers such that

$$n^{\pm}(R) \le \frac{\nu_0 R}{1 \mp \varepsilon}. \tag{A.19}$$

If $n < n^-(R)$, (A.35) shows that there exists some positive constants C and c such that

$$\left| d_r \eta_\nu^{(n,0)}(R) - i\nu_0 \eta_\nu^{(n,0)}(R) \right| \le \frac{C}{R^{3/4}} \left(\frac{n}{c} \right)^{-n}, \qquad c < n < n^-(R),$$

from which we deduce that $B_1(R) \to 0$ when $R \to +\infty$. If $n^-(r) \le n \le n^+(R)$, the estimates given in (A.33) yield

$$\left| d_r \eta_\nu^{(n,0)}(R) - i\nu_0 \eta_\nu^{(n,0)}(R) \right| \le C \left(\frac{n}{c} \right)^{-n},$$

whence we infer that

$$B_2(R) \le C \sum_{n \ge n^-(R)} n \left(\frac{n}{c} \right)^{-2n},$$

since $R \le n(1+\varepsilon)/\nu_0$. The convergence of $B_2(R)$ to 0 follows. Finally, if $n > n^+(R)$, we see from (A.29) and (A.31) that

$$\left| d_r \eta_\nu^{(n,0)}(R) - i\nu_0 \eta_\nu^{(n,0)}(R) \right| \le \frac{Cn}{R} \left(\frac{R}{c} \right)^{-n},$$

from which we easily deduce that $B_3(R) \to 0$. Property (A.17) is thus proved.

A.3. Some properties of the modified Bessel functions. We prove here some asymptotic properties of K_n and K_n' which are used in the two preceding paragraphs. The main tools are the following uniform asymptotic expansions (see Olver [24, Chap. 10, §§7 and 8]):

$$K_n(nz) = \left(\frac{\pi}{2n} \right)^{1/2} \frac{e^{-n\xi(z)}}{(1+z^2)^{1/4}} (1 + \varepsilon_0(z,n)), \tag{A.20}$$

$$K_n'(nz) = -\left(\frac{\pi}{2n} \right)^{1/2} \frac{(1+z^2)^{1/4} e^{-n\xi(z)}}{z} (1 + \varepsilon_1(z,n)), \tag{A.21}$$

where

$$\xi(z) = (1+z^2)^{1/2} + \ln \frac{z}{1 + (1+z^2)^{1/2}}. \tag{A.22}$$

These expansions are valid for $n > 0$ and z in a complex domain such as the one shown in Fig. A.1, which excludes in particular the singular points $\pm i$ of $\xi(z)$. In this domain, $\xi(z)$ is an analytic function, the branch being chosen such that $\xi(z)$ takes its principal value when $z \in \mathbb{R}^+$. Outside a vicinity of the singular points $\pm i$, the bounds for the error terms ε_0 and ε_1 are given by

$$|\varepsilon_k(z,n)| \le C \frac{1}{n \mid 1 + z^2 \mid^{1/2}}, \qquad k = 0, 1. \tag{A.23}$$

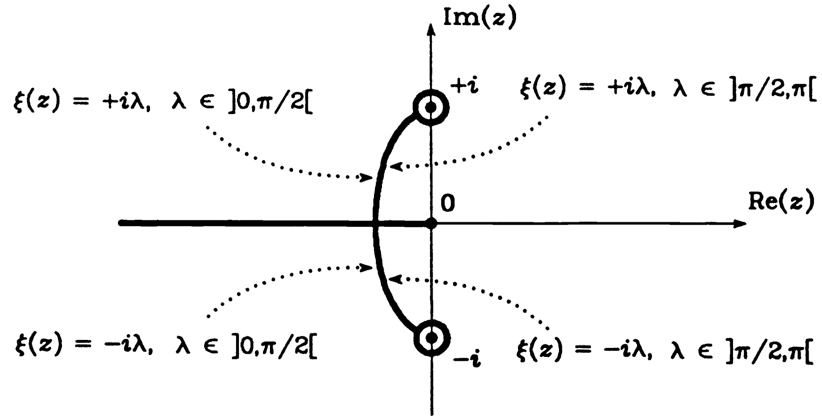


FIG. A.1. *Validity domain of the expansions of K_n and K_n'.*

We begin with the case of real positive arguments.

LEMMA A.1. *Let x_0 be a given real positive number. Then, there exist some positive constants (all denoted by C) such that*

$$\left|\frac{K_n'(x)}{K_n(x)}\right| \le C\frac{(n^2+x^2)^{1/2}}{x} \quad \textit{and} \quad \left|\frac{K_n(x)}{K_n(x_0)}\right| \le C \quad \textit{for } n \ge 0, \quad x \ge x_0. \tag{A.24}$$

Moreover, we have the more precise estimates

$$\left|\frac{K_n(x)}{K_n(x_0)}\right| \le Ce^{-(x-x_0)/3} \quad \textit{for } 0 < x_0 < x \quad \textit{and} \quad 0 \le n \le x, \tag{A.25}$$

$$\left|\frac{K_n(x)}{K_n(x_0)}\right| \le C\left(\frac{x}{2x_0}\right)^{-n} \quad \textit{for } 0 < 2x_0 < x < n. \tag{A.26}$$

Proof. First, assume that $n > 0$. Setting $z = x/n$ in (A.20) and (A.21), we derive

$$\frac{K_n'(x)}{K_n(x)} = -\frac{(n^2+x^2)^{1/2}}{x}\left(1 + O\left((n^2+x^2)^{-1/2}\right)\right),$$

which implies the first inequality in (A.24). On the other hand, we have

$$\frac{K_n(x)}{K_n(x_0)} = \left(\frac{1+x_0^2n^{-2}}{1+x^2n^{-2}}\right)^{1/4} e^{-a_n(x_0,x)}\left(1 + O\left((n^2+x_0^2)^{-1/2}\right)\right), \tag{A.27}$$

where

$$\begin{aligned} a_n(x_0,x) &= n\left(\xi(x/n) - \xi(x_0/n)\right), \\ &= \frac{x^2-x_0^2}{(n^2+x^2)^{1/2}+(n^2+x_0^2)^{1/2}} + n\ln\left(\frac{x_0^{-1}+(n^{-2}+x_0^{-2})^{1/2}}{x^{-1}+(n^{-2}+x^{-2})^{1/2}}\right). \end{aligned}$$

Since $x \ge x_0$, this latter quantity is the sum of two positive terms: the second inequality in (A.24) follows. If $n \le x$, by neglecting the second term of the right-hand side, we easily obtain

$$a_n(x_0,x) \ge (x-x_0)/3,$$

from which we deduce (A.25). If $n > x$, we neglect the first term in the expression of $a_n(x_0, x)$, which yields

$$a_n(x_0, x) \geq n \ln \frac{x}{2x_0}.$$

Inequality (A.26) follows.

Finally, if $n = 0$, we simply use the following asymptotic expansions of $K_0(x)$ and $K_0'(x)$ for large $x > 0$ (see [1]):

$$K_0(x) = \sqrt{\frac{\pi}{2x}} e^{-x}(1 + O(x^{-1})) \quad \text{and}$$
$$K_0'(x) = -K_1(x) = -\sqrt{\frac{\pi}{2x}} e^{-x}(1 + O(x^{-1})),$$

which shows that (A.24) and (A.25) hold for $n = 0$. □

We now deal with the case of imaginary arguments. The results depend on the relative positions of the order n and the modulus of the argument iy (y real): in the three following lemmas, we distinguish the three cases

$$\text{(A.28)} \qquad \frac{|y|}{n} \leq 1 - \varepsilon, \quad \frac{|y|}{n} \geq 1 - \varepsilon \quad \text{and} \quad \frac{|y|}{n} \geq 1 + \varepsilon,$$

where $\varepsilon > 0$ is an arbitrary small (but fixed) constant. All proofs are given for positive imaginary parts: the case of negative imaginary parts follows from the symmetry property

$$K_n(\bar{z}) = \overline{K_n(z)}, \qquad z \in \mathbb{C}.$$

LEMMA A.2. *Let y_0 and y_1 be two given real numbers such that $0 < |y_0| < |y_1|$. Then, there exist some positive constants (denoted by C or c) such that*

$$\text{(A.29)} \qquad \left|\frac{K_n'(iy)}{K_n(iy)}\right| \leq C \frac{n}{|y|} \quad \textit{for } 0 < |y| \leq (1-\varepsilon)n,$$

$$\text{(A.30)} \qquad \left|\frac{K_n(iy)}{K_n(iy_0)}\right| \leq C \quad \textit{for } |y_0| \leq |y| \leq |y_1| \leq (1-\varepsilon)n,$$

$$\text{(A.31)} \qquad \left|\frac{K_n(iy)}{K_n(iy_0)}\right| \leq C \left(\frac{|y|}{c}\right)^{-n} \quad \textit{for } c < |y| < (1-\varepsilon)n.$$

Proof. Setting $z = iy/n$, we see that z belongs to the validity domain of (A.20) and (A.21) shown in Fig. A.1; consequently,

$$\frac{K_n'(iy)}{K_n(iy)} = -\left(\frac{n^2}{y^2} - 1\right)^{1/2} \left(1 + O\left((n^2 - y^2)^{-1/2}\right)\right).$$

Since $n - y \geq \varepsilon n$, the error term is bounded: relation (A.29) follows. On the other hand, we have, instead of (A.27),

$$\text{(A.32)} \qquad \frac{K_n(iy)}{K_n(iy_0)} = \left(\frac{n^2 - y_0^2}{n^2 - y^2}\right)^{1/4} e^{-a_n(iy_0, iy)} \left(1 + O\left((n^2 - y^2)^{-1/2}\right)\right),$$

where

$$a_n(iy_0, iy) = \frac{-(y^2 - y_0^2)}{(n^2 - y^2)^{1/2} + (n^2 - y_0^2)^{1/2}} + n \ln \left(\frac{y_0^{-1} + (y_0^{-2} - n^{-2})^{1/2}}{y^{-1} + (y^{-2} - n^{-2})^{1/2}}\right).$$

The first term in the expression of $a_n(iy_0, iy)$ is negative, but the second one is positive and becomes dominant for large n and y bounded: inequality (A.30) is thus proved.

We obtain the more precise estimate (A.31) by noticing that, if $y \leq (1-\varepsilon)n$,

$$\frac{y^2 - y_0^2}{(n^2-y^2)^{1/2} + (n^2-y_0^2)^{1/2}} \leq Cn \quad \text{and} \quad \frac{y_0^{-1} + (y_0^{-2} - n^{-2})^{1/2}}{y^{-1} + (y^{-2}-n^{-2})^{1/2}} \geq \frac{y}{2y_0},$$

from which we deduce that, for some $c > 0$, $a_n(iy_0, iy) \geq n \ln(y/c)$ when $y > c$. □

LEMMA A.3. *Let $y_0 \neq 0$ be a given real number. There exist some positive constants (denoted by C or c) such that*

$$\text{(A.33)} \qquad \left|\frac{K_n(iy)}{K_n(iy_0)}\right| \leq C\left(\frac{n}{c}\right)^{-n} \quad \textit{and} \quad \left|\frac{K_n'(iy)}{K_n(iy_0)}\right| \leq C\left(\frac{n}{c}\right)^{-n} \quad \textit{for } c < n \leq \frac{|y|}{1-\varepsilon}.$$

Proof. In the "transition region" $y/n \in [1-\varepsilon, 1+\varepsilon]$, we cannot use the asymptotic expansions (A.20) and (A.21). However, we have the following estimate (Olver [24, Chap. 7, ex. 13.4]):

$$|K_n(iy)| \leq \left(\frac{\pi}{2y}\right)^{1/2} e^{n^2/y} \quad \text{for } n \geq 1, \quad y > 0,$$

and thus

$$|K_n(iy)| \leq \left(\frac{\pi}{2y}\right)^{1/2} e^{n/(1-\varepsilon)} \quad \text{for } y > (1-\varepsilon)n.$$

Furthermore, from the asymptotic behaviour of $K_n(iy_0)$ for fixed $y_0 > 0$ (see [1]),

$$\text{(A.34)} \qquad |K_n(iy_0)| \sim \left(\frac{\pi}{2n}\right)^{1/2} \left(\frac{2n}{ey_0}\right)^n \quad \text{when } n \to +\infty,$$

we infer that for large enough n and $y > (1-\varepsilon)n$,

$$\left|\frac{K_n(iy)}{K_n(iy_0)}\right| \leq C\, e^{n(1-\varepsilon-\ln(2n/ey_0))},$$

which implies the first inequality in (A.33). Finally, from the recurrence relation

$$K_n'(z) = -\tfrac{1}{2}\left(K_{n+1}(z) + K_{n-1}(z)\right), \qquad z \in \mathbb{C},$$

we deduce in the same way the second inequality of (A.33). □

LEMMA A.4. *Let $y_0 \neq 0$ be a given real number. There exist some positive constants C and c such that*

$$\text{(A.35)} \qquad \left|\frac{K_n'(iy) + K_n(iy)}{K_n(iy_0)}\right| \leq \frac{C}{|y|^{3/4}}\left(\frac{n}{c}\right)^{-n} \quad \textit{for } c < n \leq \frac{|y|}{1+\varepsilon}.$$

Proof. Setting $z = iy/n$, we clearly have $|z| \geq 1+\varepsilon$, and thus we can use the asymptotic expansions (A.20) and (A.21) with the error bounds (A.23). Hence

$$K_n(iy) = \left(\frac{\pi}{2}\right)^{1/2} \frac{e^{-n\xi(iy/n) - i\pi/4}}{(y^2-n^2)^{1/4}} \left(1 + O\left((y^2-n^2)^{-1/2}\right)\right),$$

$$K_n'(iy) = -\left(\frac{\pi}{2}\right)^{1/2} \frac{(y^2-n^2)^{1/4} e^{-n\xi(iy/n) - i\pi/4}}{y} \left(1 + O\left((y^2-n^2)^{-1/2}\right)\right).$$

From (A.22), we easily verify that for $y/n > 1$,

$$\xi(iy/n) = i\left((y^2n^{-2} - 1)^{1/2} + \arcsin\frac{n}{y}\right),$$

which is purely imaginary, and thus

$$K_n(iy) = \left(\frac{\pi}{2}\right)^{1/2} \frac{e^{-n\xi(iy/n)-i\pi/4}}{(y^2-n^2)^{1/4}} + O\left((y^2-n^2)^{-3/4}\right),$$
$$K_n'(iy) = -\left(\frac{\pi}{2}\right)^{1/2} \frac{(y^2-n^2)^{1/4}e^{-n\xi(iy/n)-i\pi/4}}{y} + O\left(y^{-1}(y^2-n^2)^{-1/4}\right).$$

Noticing that $y^2 - n^2 \geq \varepsilon\, ny$, we see that the error terms is these expressions are bounded by $C\,(ny)^{-3/4}$; then

$$|K_n'(iy) + K_n(iy)| = \left(\frac{\pi}{2}\right)^{1/2} \frac{n^2}{y\,(y^2-n^2)^{1/4}\left(y+(y^2-n^2)^{1/2}\right)} + O((ny)^{-3/4}),$$

which shows that

$$|K_n'(iy) + K_n(iy)| \leq C\; n^{1/4}y^{-3/4}.$$

As in the proof of Lemma A.3, using the asymptotic behaviour (A.34) of $K_n(iy_0)$, we thus deduce (A.35). □

Appendix B. Solution of the dispersion equation in the complex plane. The purpose of this section is to show a method which provides an explicit form of the roots $\zeta \in \mathbb{C}$ of the "dispersion equation" (see (3.4))

$$f(\zeta) = \nu, \quad \text{where } f(\zeta) = -\zeta\tan\zeta \tag{B.1}$$

for any complex number ν. This method, which is based on the properties of Cauchy integrals on arcs, is widely described in Henrici [7, Vol. III, Chap. 14]: some results will be stated here without proof.

B.1. Some general properties of the roots. Let ν_0 be a given complex number; suppose that (B.1) admits a solution $\zeta_0 \in \mathbb{C}$ for $\nu = \nu_0$. Function f is analytic in a vicinity of ζ_0 and can thus be represented by a series

$$f(\zeta) = \nu_0 + \sum_{m\geq 1} a_m(\zeta - \zeta_0)^m.$$

Let $m_0 \geq 1$ be the smallest value of m such that $a_m \neq 0$. We know (see, e.g., Henrici [7, Thm. 2.4f]) that in a vicinity of ν_0 (B.1) admits exactly m_0 solutions which tend to ζ_0 when $\nu \to \nu_0$: these are the m_0 branches of an analytic function having an algebraic singularity of order $m_0 - 1$ at point ν_0 (Knopp [14]). Since $a_m = d_\zeta^m f(\zeta_0)/m!$, we have

$$a_1 = -\frac{\sin 2\zeta_0 + 2\zeta_0}{2\cos^2\zeta_0} \quad \text{and} \quad a_2 = a_1\tan\zeta_0 - 1. \tag{B.2}$$

The first coefficient a_1 vanishes if $\sin 2\zeta_0 + 2\zeta_0 = 0$. In this case, $a_2 \neq 0$, which shows that ζ_0 is a double root of (B.1).

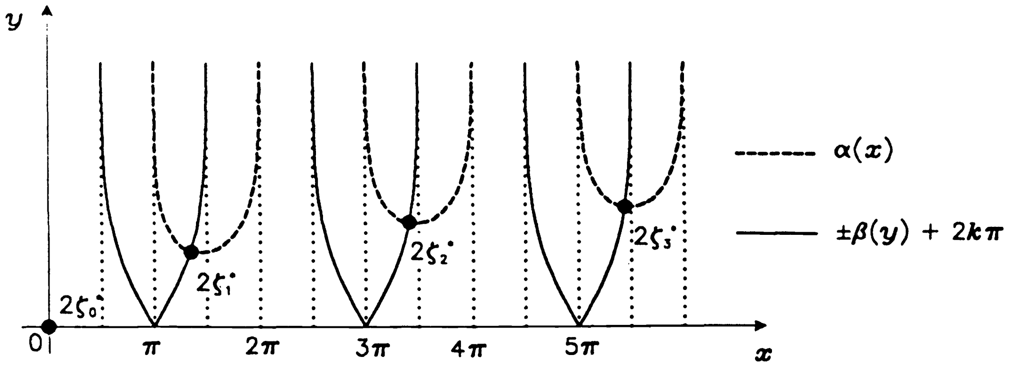


FIG. B.1. *Solutions of equation* $\sin 2\zeta + 2\zeta = 0$.

LEMMA B.1. *The roots of the following equation*

$$\sin 2\zeta + 2\zeta = 0, \tag{B.3}$$

which are such that $\operatorname{Re}\zeta \geq 0$ *and* $\operatorname{Im}\zeta \geq 0$ *form a sequence* (ζ_n^*; $n \in \mathbb{N}$)*, where* $\zeta_0^* = 0$ *and* $\operatorname{Re}\zeta_n^* \in](n-1/2)\pi, (n-1/4)\pi[$. *Their asymptotic behaviour is given by*

$$\zeta_n^* = (n-1/4)\pi + \frac{i}{2}\operatorname{Log}((4n-1)\pi) + o(1) \quad \textit{when } n \to +\infty. \tag{B.4}$$

The other roots of (B.3) *are obtained by symmetry with respect to the real and imaginary axes.*

Proof. First, notice that if ζ is a solution of (B.3), then $-\zeta$ and $\bar{\zeta}$ are ones too; moreover, $\zeta = 0$ is the only solution located on the axes $\operatorname{Re}\zeta = 0$ and $\operatorname{Im}\zeta = 0$. Consequently, setting $2\zeta = x + iy$ (x and y real), we can restrict ourselves to the case $x > 0$ and $y > 0$. Equation (B.3) is thus equivalent to

$$y = \alpha(x) = \operatorname{Arccosh}\frac{-x}{\sin x} \quad \text{and} \tag{B.5}$$

$$x = \pm\beta(y) + 2n\pi = \pm\operatorname{Arccos}\frac{-y}{\sinh y} + 2n\pi, \qquad n \in \mathbb{N}. \tag{B.6}$$

These functions are represented in Fig. B.1: the two families of curves intersect at one point $x_n^* + iy_n^* = 2\zeta_n^*$ in every strip $x \in](2n-1)\pi, (2n-1/2)\pi[$. It may be easily seen that $\operatorname{Min}\alpha(x)$ on each interval $]n\pi, (n+1)\pi[$ tends to $+\infty$ when $n \to +\infty$, when we infer that

$$x_n^* = (2n-1/2)\pi + o(1) \quad \text{when } n \to +\infty.$$

Substituting this expression in (B.5) yields

$$y_n^* = \operatorname{Log}((4n-1)\pi) + o(1).$$

The asymptotic behaviour (B.4) of ζ_n^* follows. □

Let $\mathbb{E}$ denote the set of ν which correspond to the solutions ζ of (B.3), i.e.,

$$\nu = -\zeta\tan\zeta = \sin^2\zeta. \tag{B.7}$$

In other words, $\nu \in \mathbb{E}$ means that (B.1) admits a double root ζ (which is a solution of (B.3)). Note that ζ and $-\zeta$ are the only double roots of (B.1) for such a ν : indeed, it is a easy matter to verify that

$$f(\zeta_1) = f(z_2) \quad \text{and} \quad \sin 2\zeta_1 + 2\zeta_1 = \sin 2\zeta_2 + 2\zeta_2 = 0$$

implies $\zeta_1 = \pm\zeta_2$. The locations of the points of $\mathbb{E}$ are represented in Fig. 3.2: $\mathbb{E}$ is symmetrical with respect to the real axis. Moreover, we have from (B.4)

$$\text{(B.8)} \qquad \nu_n^* = f(\zeta_n^*) = -i\pi(n - 1/4)(1 + o(1)) \quad \text{when } n \to +\infty.$$

These results may be summarized as follows.

PROPOSITION B.1. *rm (i) If $\nu \in \mathbb{C}\backslash\mathbb{E}$, then every root of the dispersion equation* (B.1) *is simple. If $\nu \in \mathbb{E}$,* (B.1) *has exactly two double roots ζ and $-\zeta$, solutions of* (B.3); *the others are simple.*

(ii) *If ζ_0 is a simple root of* (B.1) *for $\nu = \nu_0$, then there exists a vicinity of ν_0 in which* (B.1) *has only one root ζ_ν which tends to ζ_0 when $\nu \to \nu_0$: this root is simple and depends analytically on ν.*

(iii) *If ζ_0 is a double root of* (B.1) *for $\nu = \nu_0$, then there exists a vicinity of ν_0 in which* (B.1) *has exactly two simple roots ζ_ν^+ and ζ_ν^-, for $\nu \neq \nu_0$, which tend to ζ_0 when $\nu \to \nu_0$: these are the two branches of an analytic function of ν which has an algebraic singularity of order* 1 *at point ν_0.*

B.2. Other formulation of the dispersion equation. Let us first note some symmetry properties. For a given ν, if ζ is a solution of (B.1), then $-\zeta$ is also a solution for the same ν and $\bar{\zeta}$ is a solution of (B.1) for $\bar{\nu}$. We can thus restrict ourselves to the case

$$\text{(B.9)} \qquad \operatorname{Re}\zeta \geq 0 \quad \text{and} \quad \operatorname{Im}\nu \geq 0.$$

Let $\operatorname{Arctan} z$ denote the principal value of the inverse of $\tan z$, i.e.,

$$\operatorname{Arctan} z = \int_0^z \frac{dt}{1+t^2},$$

where the path of integration must not cross the two cuts $\{iy; y \in \mathbb{R} \text{ and } |y| \geq 1\}$. If ν/ζ is assumed not to belong to one of these cuts, (B.1) (with the restriction (B.9)) is equivalent to

$$\text{(B.10)} \qquad \zeta = -\operatorname{Arctan}(\nu/\zeta) + k\pi, \qquad k \geq 0.$$

Note that for a given k, the roots of (B.10), if any, are such that

$$\text{(B.11)} \qquad \operatorname{Re}\zeta \in \left](k - \tfrac{1}{2})\pi, \quad (k + \tfrac{1}{2})\pi\right[.$$

It may happen that ν/ζ belongs to one of the cuts of function $\operatorname{Arctan} z$. Indeed, suppose that $\zeta = -i\nu t$ is a solution of (B.1), where t is a real number such that $0 < |t| \leq 1$. From the symmetry properties stated above, we can assume $t > 0$. Equation (B.1) is thus equivalent to $t\tanh(\nu t) = 1$, which yields

$$\text{(B.12)} \qquad \nu = \frac{1}{t}\left[\frac{1}{2}\operatorname{Log}\frac{1+t}{1-t} + i\pi\left(n + \frac{1}{2}\right)\right], \qquad t \in]0,1[, \quad n \geq 0.$$

This expression defines a family of curves ($C_n \subset \mathbb{C}$; $n \geq 0$) which are represented in Fig. B.2.

For every $n > 0$, we denote by D_n the open set located between C_{n-1} and C_n ; D_0 is the domain of $\mathbb{C}^+ \setminus \{0\}$ situated below C_0. We define D and C by

$$\text{(B.13)} \qquad D = \bigcup_{n\geq 0} D_n \quad \text{and} \quad C = \bigcup_{n\geq 0} C_n.$$

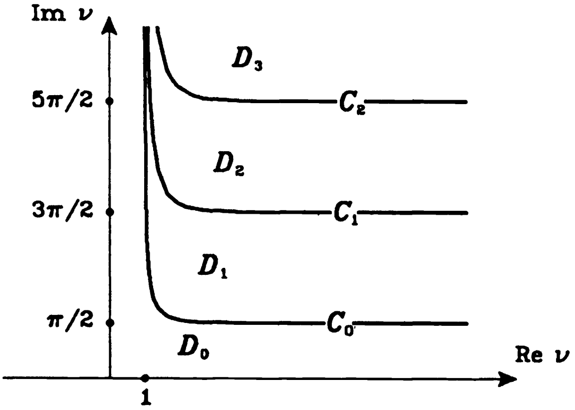


FIG. B.2. *Definition of D_n and C_n.*

In the following paragraph, we will assume that $\nu \in D$. The case $\nu \in C$ will be dealt with at the end of §B.4. Note that this distinction is not specific to (B.1): it results from the choice of the cuts which are necessary for the definition of function $\operatorname{Arctan} z$.

B.3. The Privalov problem. Using the change of variable

$$x = \zeta/(i\nu); \tag{B.14}$$

(B.10) amounts to determining the zeros of functions $f_\nu^{(k)}(x)$ defined as follows

$$f_\nu^{(k)}(x) = \nu x + ik\pi + \frac{1}{2}\operatorname{Log}\frac{x-1}{x+1}, \tag{B.15}$$

where Log is the principal value of the logarithm, and k is a nonnegative integer. Let I denote the segment $[-1,+1]$ oriented from -1 to $+1$. For every $\nu \in D$ and every $k \geq 0$, function $f_\nu^{(k)}$ is analytic on $\mathbb{C} \setminus I$. Let t be a point on I different from the endpoints ± 1 : such a point will be called an interior point of I. If $B(t)$ is a sufficiently small disk centered at t, the set $B(t) \setminus I$ has two separated components

$$B^+(t) = B(t) \cap \{x \in \mathbb{C};\ \operatorname{Im} x > 0\},$$
$$B^-(t) = B(t) \cap \{x \in \mathbb{C};\ \operatorname{Im} x < 0\},$$

and we can define the "one-side limits" of $f_\nu^{(k)}$ at t by

$$f_\nu^{(k)\pm}(t) = \lim_{x \in B^\pm(t) \to t} f_\nu^{(k)}(x).$$

It may be easily verified that at every interior point t of I these limits exist and read

$$f_\nu^{(k)\pm}(t) = t\nu + i\pi(k \pm 1/2) + \frac{1}{2}\operatorname{Log}\frac{1-t}{1+t}. \tag{B.16}$$

Note that the assumption $\nu \in D$ implies that $f_\nu^{(k)+}$ and $f_\nu^{(k)-}$ do not vanish at any interior point of I.

Consider then the function $a_\nu^{(k)}(t)$ defined for every interior point of I by

$$a_\nu^{(k)}(t) = f_\nu^{(k)+}(t)\ \left(f_\nu^{(k)-}(t)\right)^{-1}. \tag{B.17}$$

The so-called (homogeneous) Privalov problem associated with $a_\nu^{(k)}$ consists in finding a function $F_\nu^{(k)}(x)$ which is analytic in $\mathbb{C}\setminus I$, has at most a pole at ∞, and has one-side limits $F_\nu^{(k)+}$ and $F_\nu^{(k)-}$ (at every interior point of I) which satisfy the relation

$$a_\nu^{(k)}(t) = F_\nu^{(k)+}(t)\ \left(F_\nu^{(k)-}(t)\right)^{-1}.$$

Function $f_\nu^{(k)}(x)$ is a particular solution of the Privalov problem. Indeed, it has a pole of order 1 at ∞, since

$$f_\nu^{(k)}(x) = \nu x + ik\pi - x^{-1} + O(x^{-2}). \tag{B.18}$$

The following result provides a general expression for any solution of the Privalov problem (see Henrici [7, §14.8]).

PROPOSITION B.2. *Let $\nu \in D$; all the solutions of the homogeneous Privalov problem are given by*

$$F_\nu^{(k)}(x) = r_\nu^{(k)}(x)\ q_\nu^{(k)}(x), \quad \textit{where} \tag{B.19}$$

$$q_\nu^{(k)}(x) = \exp\left(\frac{1}{2i\pi}\int_I \frac{\log a_\nu^{(k)}(t)}{t-x}\,dt\right), \tag{B.20}$$

(*where* log *denotes any continuous logarithm*) *and $r_\nu^{(k)}(x)$ is an analytic function in $\mathbb{C}\setminus\{\pm 1\}$ which has at most a pole at ∞.*

Proof. First, consider the expression (B.20) of $q_\nu^{(k)}$. Since $f_\nu^{(k)+}$ and $f_\nu^{(k)-}$ are differentiable functions and do not vanish at any interior point of I, the same holds for $a_\nu^{(k)}$. Moreover,

$$\lim_{t\to\pm 1} a_\nu^{(k)}(t) = 1, \tag{B.21}$$

which shows that $a_\nu^{(k)}$ does not vanish on I. For any continuous logarithm, $\log a_\nu^{(k)}$ is a continuous function on I, and we can then define the Cauchy integral

$$\ell_\nu^{(k)}(x) = \frac{1}{2\pi i}\int_I \frac{\log a_\nu^{(k)}(t)}{t-x}\,dt, \qquad x \in \mathbb{C}\setminus I. \tag{B.22}$$

Function $\log a_\nu^{(k)}$ is differentiable on every closed segment of I which does not contain ± 1. Consequently, the one-side limits $\ell_\nu^{(k)+}(t)$ et $\ell_\nu^{(k)-}(t)$ exist at every interior point of I and satisfy the following relation (Sokhotskyi formulas; see Henrici [7, Thm. 14.1c]):

$$\ell_\nu^{(k)+}(t) = \ell_\nu^{(k)-}(t) + \log a_\nu^{(k)}(t).$$

Thus, function $q_\nu^{(k)}(x)$ given by (B.20) is a particular solution of the Privalov problem (it is obviously analytic at ∞), and every function of the form (B.19) is also a solution.

Conversely, let $F_\nu^{(k)}(x)$ denote any solution of the Privalov problem. Since $q_\nu^{(k)}$ vanishes neither on $\mathbb{C}\setminus I$ nor at ∞, function

$$r_\nu^{(k)}(x) = F_\nu^{(k)}(x)\left(q_\nu^{(k)}(x)\right)^{-1}$$

is analytic in $\mathbb{C} \setminus I$ and has at most a pole at ∞. Moreover, $r_\nu^{(k)+}(t) = r_\nu^{(k)-}(t)$ at every interior point of I : $r_\nu^{(k)}$ is thus analytic in $\mathbb{C} \setminus \{\pm 1\}$, which completes the proof. □

Proposition B.2 shows that $f_\nu^{(k)}$ can be written in the form (B.19). From now on, $r_\nu^{(k)}$ will denote the following function:

$$\text{(B.23)} \qquad r_\nu^{(k)}(x) = f_\nu^{(k)}(x)\ \left(q_\nu^{(k)}(x)\right)^{-1} = f_\nu^{(k)}(x)\ \exp\left(-\ell_\nu^{(k)}(x)\right),$$

whose zeros coincide with the ones of $f_\nu^{(k)}$ (if they differ from ± 1). By studying its behaviour in the vicinity of the endpoints ± 1, as well as at ∞, we prove below that $r_\nu^{(k)}$ actually is a rational function.

We first have to precise the definition of $\log a_\nu^{(k)}$ which was so far defined up to a multiple of $2i\pi$. From (B.21), we can choose, for instance,

$$\text{(B.24)} \qquad \lim_{t \to -1} \log a_\nu^{(k)}(t) = 0,$$

which selects a particular branch of the logarithm. According to this choice, we have the following result.

LEMMA B.2. *When $t \to +1$, the limit of $\log a_\nu^{(k)}(t)$ is given by*

$$\text{(B.25)} \qquad \lim_{t \to +1} \log a_\nu^{(k)}(t) = \begin{cases} -2i\pi & \textit{if } k = 0 \quad \textit{and} \quad \nu \notin D_0, \\ 0 & \textit{if } k \neq 0 \quad \textit{and} \quad \nu \notin D_k, \\ +2i\pi & \textit{if } \nu \in D_k. \end{cases}$$

Proof. $a_\nu^{(k)}(I)$ is a curve of the complex plane which is continuous, closed (from (B.21)), and does not pass through 0. We thus have to determine the index of this curve with respect to 0 :

$$\mathcal{I}_\nu^{(k)} = \frac{1}{2i\pi} \int_{a_\nu^{(k)}(I)} \frac{dx}{x} = \frac{1}{2i\pi} \int_{-1}^{+1} d\left[\log a_\nu^{(k)}(t)\right].$$

It may be easily verified that for a given $k \geq 0$, $\mathcal{I}_\nu^{(k)}$ is a continuous (in fact, analytic) function of ν in every domain D_n. Since it is always an integer, we deduce that it is constant in D_n. In order to calculate it, we study below the shape of $a_\nu^{(k)}(I)$ for the particular point $\nu = \varepsilon^{-1} + in\pi$ of D_n, where ε is an arbitrary small positive number. We have, from (B.17),

$$\operatorname{Re} a_\nu^{(k)}(t) = \frac{(g_\varepsilon(t))^2 + ((tn+k)^2 - \frac{1}{4})\pi^2}{(g_\varepsilon(t))^2 + (tn+k-\frac{1}{2})^2\pi^2},$$

$$\operatorname{Im} a_\nu^{(k)}(t) = \frac{\pi\, g_\varepsilon(t)}{(g_\varepsilon(t))^2 + (tn+k-\frac{1}{2})^2\pi^2},$$

where

$$g_\varepsilon(t) = t\varepsilon^{-1} + \frac{1}{2} \operatorname{Log} \frac{1-t}{1+t}.$$

By studying the variations of this function, we see that the sign of $\operatorname{Im} a_\nu^{(k)}(t)$ is given by

t	-1		$-t_\varepsilon^*$		0		$+t_\varepsilon^*$		$+1$
$\operatorname{Im} a_\nu^{(k)}(t)$	0	+	0	−	0	+	0	−	0

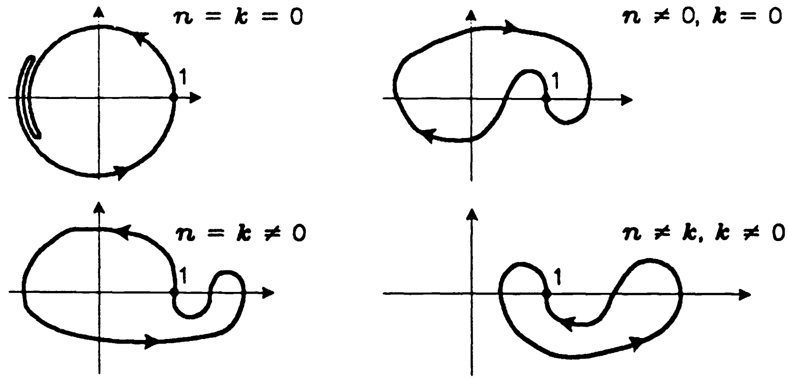


FIG. B.3. *Representation of* $a_\nu^{(k)}(I)$.

where t_ε^* is a real number such that $\sqrt{1-\varepsilon} < t_\varepsilon^* < 1$. Therefore, the curve $a_\nu^{(k)}(I)$ crosses successively the real axis at the four points

$$1 = a_\nu^{(k)}(\pm 1), \quad x_\varepsilon^- = a_\nu^{(k)}(-t_\varepsilon^*), \quad \frac{k+1/2}{k-1/2} = a_\nu^{(k)}(0) \text{ and } x_\varepsilon^+ = a_\nu^{(k)}(+t_\varepsilon^*).$$

Noticing that t_ε^* can be chosen arbitrarily close to 1, we infer that

$$x_\varepsilon^\pm \sim \frac{\pm n + k + 1/2}{\pm n + k - 1/2} \quad \text{when } \varepsilon \to 0.$$

We thus deduce the shape of $a_\nu^{(k)}(I)$ according to k and n (see Fig. B.3), which shows that

$$\mathcal{I}_\nu^{(k)} = \begin{cases} -1 & \text{if } k = 0 \text{ and } \nu \notin D_0, \\ 0 & \text{if } k \neq 0 \text{ and } \nu \notin D_k, \\ +1 & \text{if } \nu \in D_k. \end{cases}$$

Consequently, the variation of the argument of $a_\nu^{(k)}(t)$ when t moves along I is -2π in the first case, 0 in the second one, and $+2\pi$ in the last one. The statement of the lemma follows. □

PROPOSITION B.3. *Function* $r_\nu^{(k)}$ *given by* (B.23) *is a rational function which has a pole of order* 1 *at* ∞ *and a removable singularity at* -1. *At point* $+1$, *the singularity is a pole of order* 1 *if* $\nu \in D_k$, *and is removable if* $\nu \in D \setminus D_k$.

Proof. Using the same techniques as the ones described by Henrici [7, §§14.7 and 14.8], we deduce from (B.24) and (B.25) that $r_\nu^{(k)}(x) = O(1)$ in the vicinity of $x = -1$, and

$$r_\nu^{(k)}(x) = \begin{cases} O((x-1)^{-1}) & \text{if } \nu \in D_k, \\ O(1) & \text{if } \nu \in D \setminus D_k, \end{cases}$$

in the vicinity of $x = +1$. Moreover, at ∞, the singularities of $f_\nu^{(k)}$ and $r_\nu^{(k)}$ are the same (i.e., a pole of order 1, from (B.18)), since $q_\nu^{(k)}$ is analytic. Consequently, $r_\nu^{(k)}$ has at most a pole at $+1$ and ∞, and is analytic elsewhere: it is thus a rational function. □

B.4. Explicit form of the solutions. The Laurent series of $r_\nu^{(k)}$ in the vicinity of ∞ can be constructed explicitly. From (B.22), we have

$$\ell_\nu^{(k)}(x) = \frac{-1}{2\pi i} x^{-1} \int_I \frac{\log a_\nu^{(k)}(t)}{1 - tx^{-1}} dt = -\sum_{p=0}^{\infty} m_{\nu,p}^{(k)} \, x^{-p-1},$$

where

$$m_{\nu,p}^{(k)} = \frac{1}{2\pi i}\int_I t^p \log a_\nu^{(k)}(t)\,dt, \qquad p \geq 0. \tag{B.26}$$

Consequently,

$$\exp(-\ell_\nu^{(k)}(x)) = 1 + \left[m_{\nu,0}^{(k)}\right]x^{-1} + \left[m_{\nu,1}^{(k)} + \tfrac{1}{2}m_{\nu,0}^{(k)^2}\right]x^{-2} + O(x^{-3}).$$

Using the Laurent series (B.18) of $f_\nu^{(k)}(x)$ at ∞, we deduce

$$r_\nu^{(k)}(x) = \nu x + \beta_\nu^{(k)} + \gamma_\nu^{(k)}x^{-1} + O(x^{-2}), \tag{B.27}$$

where

$$\beta_\nu^{(k)} = ik\pi + \nu\, m_{\nu,0}^{(k)}, \tag{B.28}$$

$$\gamma_\nu^{(k)} = -1 + ik\pi m_{\nu,0}^{(k)} + \nu\left(m_{\nu,1}^{(k)} + \tfrac{1}{2}m_{\nu,0}^{(k)^2}\right). \tag{B.29}$$

On the other hand, Proposition B.3 shows that $r_\nu^{(k)}(x)$ has the form $ax + b + c(x-1)^{-1}$, where $c = 0$ if $\nu \in D \setminus D_k$. Comparing the Laurent series of this function with (B.27), we deduce that $a = \nu$, $b = \beta_\nu^{(k)}$, and $c = \gamma_\nu^{(k)}$. The roots of $r_\nu^{(k)}$, which coincide with the ones of $f_\nu^{(k)}$ if they differ from ± 1, are thus given by

$$\begin{aligned} x_\nu^{(k)} &= -\beta_\nu^{(k)}/\nu \quad \text{if } \nu \in D \setminus D_k, \\ x_\nu^{(k)\pm} &= \left(\nu - \beta_\nu^{(k)} \pm \sqrt{(\beta_\nu^{(k)} + \nu)^2 - 4\nu\gamma_\nu^{(k)}}\right)\Big/ 2\nu \quad \text{if } \nu \in D_k. \end{aligned}$$

In the particular case $k = 0$, it may be easily seen that $m_{\nu,0}^{(0)} = 1$, which shows that $\beta_\nu^{(0)} = \nu$. Consequently, if $\nu \in D \setminus D_0$, we have $x_\nu^{(0)} = -1$ which is not acceptable, and if $\nu \in D_0$, the two roots $x_\nu^{(0)\pm}$ are opposite to each other (see the symmetry properties in §B.2): they differ from ± 1 since 1 is a pole of $r_\nu^{(0)}$ when $\nu \in D_0$.

Finally, according to the change of variable (B.14) and property (B.11), we have the following.

PROPOSITION B.4. *The roots of the dispersion equation* (B.1) *are distributed as follows:*

(i) *if $\nu \in D_n$ for $n \neq 0$, no root in the strip $]-\pi/2, \pi/2[\, \times i\mathbb{R}$, one root $z_\nu^{(k)}$ in each strip $](k-\frac{1}{2})\pi, (k+\frac{1}{2})\pi[\, \times i\mathbb{R}$ for $k \neq n$, and two roots $z_\nu^{(n)\pm}$ in the strip $](n-\frac{1}{2})\pi, (n+\frac{1}{2})\pi[\, \times i\mathbb{R}$, where*

$$z_\nu^{(k)} = -i\beta_\nu^{(k)} = k\pi - i\nu m_{\nu,0}^{(k)}, \tag{B.30}$$

$$z_\nu^{(k)\pm} = \frac{i}{2}\left(\nu - \beta_\nu^{(k)} \pm \sqrt{(\beta_\nu^{(k)} + \nu)^2 - 4\nu\gamma_\nu^{(k)}}\right); \tag{B.31}$$

(ii) *if $\nu \in D_0$, two opposite roots $z_\nu^{(0)\pm}$ in the strip $]-\pi/2, \pi/2[\, \times i\mathbb{R}$ given by* (B.31) *with $\beta_\nu^{(0)} = \nu$, and one root $z_\nu^{(k)}$ in each strip $](k-\frac{1}{2})\pi, (k+\frac{1}{2})\pi[\, \times i\mathbb{R}$ for $k \neq 0$, where $z_\nu^{(k)}$ is still given by* (B.30).

Remark B.1. The two roots located in the strip $](n-\frac{1}{2})\pi,(n+\frac{1}{2})\pi[\times i\mathbb{R}$ (for $\nu\in D_n$) may coincide: more precisely, there exists one (and only one) value of $\nu\in D_n$ such that $x_\nu^{(n)+}=x_\nu^{(n)-}$ is a double root of the dispersion equation. Indeed, we know from Lemma B.1 that there is one possible double root ζ_n^* of (B.1) in each strip $](n-\frac{1}{2})\pi,(n+\frac{1}{2})\pi[\times i\mathbb{R}$: this root corresponds to the exceptional point $\nu_n^*=f(\zeta_n^*)=\sin^2\zeta_n^*\in\mathbb{E}$. Consequently, each domain D_n contains one (and only one) point of $\mathbb{E}$.

Consider finally the case $\nu\in C$. By virtue of the definition (B.12) of the curves C_n, this means that there exists $n\geq 0$ and $t\in]0,1[$ such that

$$\nu=\frac{1}{t}\left[\frac{1}{2}\operatorname{Log}\frac{1+t}{1-t}+i\pi\left(n+\frac{1}{2}\right)\right]. \tag{B.32}$$

Moreover,

$$\zeta_{n,t}=-i\nu t=\pi\left(n+\frac{1}{2}\right)-\frac{i}{2}\operatorname{Log}\frac{1+t}{1-t} \tag{B.33}$$

is a simple root of the dispersion equation (B.1). In fact, by Remark B.1, all the roots of (B.1) are simple in this case. Proposition B.1 thus shows that there exists a vicinity of ν in which these roots are analytic. Consequently, their expressions are obtained by going to the limit ($\nu\longrightarrow\nu_n\in C_n$) in (B.30) and (B.31). For some of them, a singular integral appears. Indeed, if $\nu\in C_n$, functions $f_\nu^{(n)+}$ and $f_\nu^{(n+1)-}=f_\nu^{(n)+}$ vanish on I at point $-t$, where t is the parameter which corresponds ν in (B.32). On the other hand, $f_\nu^{(k)+}$ (respectively, $f_\nu^{(k)-}$) does not vanish on I if $k\neq n$ (respectively, $k\neq n+1$). We finally deduce from Proposition B.4 the following:

(i) if ν passes from D_0 to D_1 across C_0, the only root of (B.1) located in the strip $]0,\pi/2[\times i\mathbb{R}$ crosses the line $\operatorname{Re}\zeta=\pi/2$ at point $\zeta_{0,t}$ given in (B.33);

(ii) if ν passes from D_n to D_{n+1} ($n>0$) across C_n, one of the two roots of (B.1) located in the strip $](n-1/2)\pi,(n+1/2)\pi[\times i\mathbb{R}$ crosses the line $\operatorname{Re}\zeta=(n+1/2)\pi$ at point $\zeta_{n,t}$ given in (B.33).

B.5. Further properties of the roots. We consider in this last paragraph the roots $\zeta_\nu^{(m)}$, $m\geq 0$, of the dispersion equation defined in Lemma 3.1 for real positive ν: $\zeta_\nu^{(0)}$ is imaginary with negative imaginary part and $\zeta_\nu^{(m)}$, for $m\geq 1$, are real, positive and arranged according to increasing values. The results of §B.4 together with Proposition B.1 show that each $\zeta_\nu^{(m)}$ extends to an analytic function of ν in any simply connected domain of $\mathbb{C}\setminus\mathbb{E}$. We state below some properties of these roots.

PROPOSITION B.5. *If* $\operatorname{Im}\nu>0$, *the roots* $\zeta_\nu^{(m)}$ *of the dispersion equation* (B.1) *are such that*

$$\operatorname{Re}\zeta_\nu^{(m)}>0 \quad \textit{and} \quad \operatorname{Im}\zeta_\nu^{(m)}<0 \quad \forall m\geq 0. \tag{B.34}$$

If $\operatorname{Im}\nu<0$, *we have*

$$\begin{aligned}&\operatorname{Re}\zeta_\nu^{(0)}<0 \quad \textit{and} \quad \operatorname{Im}\zeta_\nu^{(0)}<0,\\ &\operatorname{Re}\zeta_\nu^{(m)}>0 \quad \textit{and} \quad \operatorname{Im}\zeta_\nu^{(m)}>0 \quad \forall m>0.\end{aligned} \tag{B.35}$$

Proof. First, notice that if ν lies in a simply connected domain of $\mathbb{C}\setminus\mathbb{E}$ which does not cross the real axis, then the signs of $\operatorname{Re}\zeta_\nu^{(m)}$ and $\operatorname{Im}\zeta_\nu^{(m)}$ (for a given $m\geq 0$) are constant. This follows from the fact that a root ζ of (B.1) may be located on the real or imaginary axes only if ν is real (since $\nu = -\zeta \tan \zeta$). In order to find these signs, we simply have to study $d_\nu \zeta_\nu^{(m)}$ when $\nu \in \mathbb{R}^+$. Deriving (B.1) with respect to ν yields

$$d_\nu \zeta_\nu^{(m)} = \frac{-2\cos^2(2\zeta_\nu^{(m)})}{\sin 2\zeta_\nu^{(m)} + 2\zeta_\nu^{(m)}}.$$

If $\zeta_\nu^{(m)}$ is imaginary (i.e., $m = 0$), this quantity is imaginary too, and its imaginary part is positive. Similarly, if $\zeta_\nu^{(m)}$ is real and positive, it is real and negative. (B.34) and (B.35) follow. □

PROPOSITION B.6. *The asymptotic behaviour of $\zeta_\nu^{(m)}$ when $m \to +\infty$ is given by*

$$\zeta_\nu^{(m)} = m\pi - m^{-1}\nu/\pi + O(m^{-3}). \tag{B.36}$$

Proof. We know from Proposition B.4 that for large enough m, $\zeta_\nu^{(m)}$ is nothing but the root $z_\nu^{(m)}$ given by (B.30). We thus have from (B.26),

$$\zeta_\nu^{(m)} = m\pi - \frac{\nu}{2\pi}\int_I \log a_\nu^{(m)}(t)\, dt,$$

where log is here the principal branch Log of the logarithm (see (B.24) and Fig. B.3 in the case $n \neq k$ and $k \neq 0$). Recall that $a_\nu^{(m)}(t)$ reads

$$a_\nu^{(m)}(t) = \frac{t\nu + i\pi(m+\frac{1}{2}) + \frac{1}{2}\operatorname{Log}\frac{1-t}{1+t}}{t\nu + i\pi(m-\frac{1}{2}) + \frac{1}{2}\operatorname{Log}\frac{1-t}{1+t}}.$$

For every $t \in]-1,+1[$, this quantity expands as follows:

$$a_\nu^{(m)}(t) = 1 + m^{-1} + m^{-2}\left(\frac{1}{2} - \frac{1}{i\pi}\left(\nu t + \operatorname{Log}\frac{1-t}{1+t}\right)\right) + O(m^{-3}).$$

Consequently,

$$\operatorname{Log} a_\nu^{(m)}(t) = m^{-1} - m^{-2}\left(\frac{1}{i\pi}\left(\nu t + \operatorname{Log}\frac{1-t}{1+t}\right)\right) + O(m^{-3}).$$

This expansion is valid uniformly with respect to t in every closed segment $[a,b] \subset]-1,+1[$: we can thus integrate it over I. Noticing that

$$\int_I \left(\nu t + \operatorname{Log}\frac{1-t}{1+t}\right) dt = 0,$$

the expansion (B.36) of $\zeta_\nu^{(m)}$ follows. □

We end this section by a remark about the role of the exceptional points of $\mathbb{E}$. We show in Fig. B.5 the path followed by each root $\zeta_\nu^{(m)}$ when ν describes the closed curve around the "first" exceptional point represented in Fig. B.4 (the dots in Fig. B.5 are the possible double roots of the dispersion equation, i.e., the solutions of (B.3)). More generally, it may be easily seen that when ν describes a closed path around the nth exceptional point, every root returns to its initial value, except $\zeta_\nu^{(0)}$ and $\zeta_\nu^{(n)}$ which

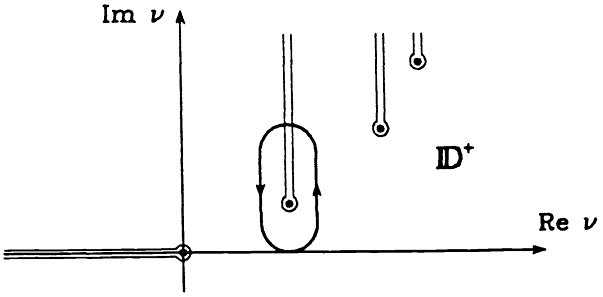


FIG. B.4. *Closed path for ν around an exceptional point.*

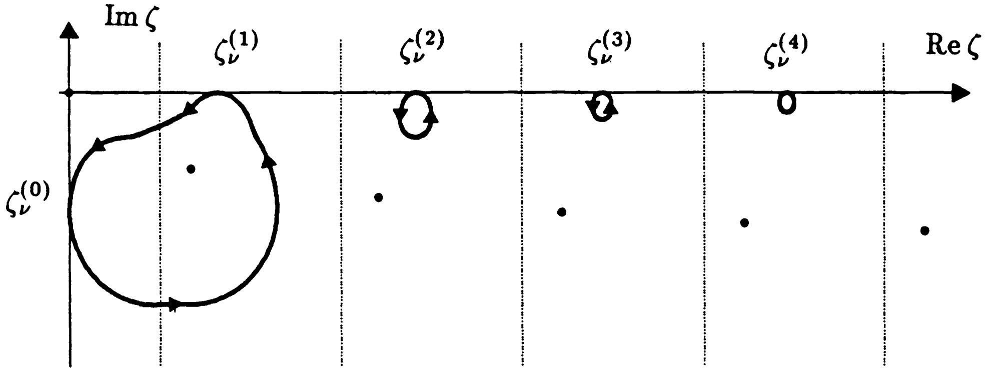


FIG. B.5. *Corresponding path of each $\zeta_\nu^{(m)}$.*

undergo a permutation. This illustrates the fact that the points of $\mathbb{E}$ are branch points of order 1.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1970.

[2] J. T. Beale, *Eigenfunction expansions for objects floating in an open sea*, Comm. Pure Appl. Math., 36 (1973), pp. 283–313.

[3] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[4] T. Hargé, *Valeurs propres d'un corps élastique*, C. R. Acad. Sci. Paris, Sér. I, 311 (1990), pp. 857–859.

[5] C. Hazard, *Etude des résonances pour le problème linéarisé des mouvements d'un navire sur la houle*, thesis, Université Pierre et Marie Curie, Paris, 1991.

[6] C. Hazard, M. Lenoir, and D. Martin, *Integral representation of resonant states for the sea-keeping problem*, manuscript.

[7] P. Henrici, *Applied and computational complex analysis*, (3 vol.), John Wiley, New York, 1974.

[8] H. Hochstadt, *Les fonctions de la physique mathématique*, Masson, Paris, 1973.

[9] A. Jami and M. Lenoir, *A variational formulation for exterior problems in linear hydrodynamics*, Comput. Methods Appl. Mech. Engrg., 16 (1978), pp. 341–359.

[10] F. John, *On the motion of floating bodies*, I, Comm. Pure Appl. Math., 2 (1949), pp. 13–57.

[11] ———, *On the motion of floating bodies*, II, Comm. Pure Appl. Math., 3 (1950), pp. 45–101.

[12] P. Joly and O. Poisson, manuscript.

[13] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1984.

[14] K. Knopp, *Theory of Functions*, Part II, English translation, Dover, New York, 1947.

[15] P. D. Lax and R. S. Phillips, *Scattering Theory*, Academic Press, New York, 1967.

[16] M. Lenoir and D. Martin, *An application of the limiting absorption principle to the motions of floating bodies*, J. Math. Anal. Appl., 79 (1981), pp. 370–383.

[17] M. Lenoir and A. Tounsi, *The localized finite element method and its application to the 2D sea-keeping problem*, SIAM J. Numer. Anal., 25 (1998), pp. 729–752.

[18] M. Lenoir, M. Vullierme-Ledard, and C. Hazard, *Variational formulations for the determination of resonant states in scattering problems*, SIAM J. Math. Anal., 23 (1992), pp. 579–608.

[19] C. Licht, *Trois modèles décrivant les vibrations d'une structure élastique dans la mer*, C. R. Acad. Sc. Paris, Sér. I, 296 (1983), pp. 341–344.

[20] J. L. Lions, *Problèmes aux limites dans les équations aux dérivées partielles*, Presses de l'Université de Montreal, Montreal, 1965.

[21] J. L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Travaux et Recherches Mathématiques, Dunod, Paris, 1968.

[22] J. Nečas and I. Hlaváček, *Mathematical Theory of Elastic and Elasto-Plastic Bodies: an Introduction*, Elsevier, Amsterdam, 1981.

[23] R. Ohayon and E. Sanchez-Palencia, *On the vibration problem for an elastic body surrounded by a slightly compressible fluid*, RAIRO Anal. Numér., 17 (1983), pp. 311–326.

[24] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[25] O. Poisson, *Calcul des pôles de résonance associés à la diffraction d'ondes acoustiques par un obstacle en dimension 2*, C. R. Acad. Sci. Paris, Sér. I, 315 (1992), pp. 747–753.

[26] ———, *Calcul des pôles de résonance associés à la diffraction d'ondes acoustiques et élastiques en dimension 2*, thesis, Université Paris IX Dauphine, Paris, 1992.

[27] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, I: *Functional Analysis*, Academic Press, New York, 1980.

[28] J.P. Richer, *Contribution à la résolution du problème de tenue à la mer aux premier et second ordres*, thesis, E.N.S.T.A., Paris, 1988.

[29] J. Sanchez-Hubert and E. Sanchez-Palencia, *Vibration and Coupling of Continuous Systems, Asymptotics Methods*, Springer-Verlag, Berlin, 1989.

[30] E. Sanchez-Palencia, *Nonhomogeneous media and vibration theory*, Springer-Verlag, Berlin, 1980.

[31] S. Steinberg, *Meromorphic families of compact operators*, Arch. Rational Mech. Anal., 31 (1968), pp. 372–380.

[32] M. Vullierme-Ledard, *Asymptotic study of the vibration problem for an elastic body deeply immersed in an incompressible fluid*, Math. Model. Numer. Anal., 19 (1985), pp. 145–170.

[33] ———, *The limiting amplitude principle applied to the motions of floating bodies*, Math. Model. Numer. Anal., 21 (1987), pp. 125–170.

[34] ———, *Problèmes asymptotiques de l'hydrodynamique navale linéarisée et du couplage fluide-structure*, thesis, Université Pierre et Marie Curie, Paris, 1987.

SIAM J. MATH. ANAL.
Vol. 24, No. 6, pp. 1515–1521, November 1993

005

GLOBAL EXISTENCE AND BLOW-UP PROBLEMS FOR QUASILINEAR PARABOLIC EQUATIONS WITH NONLINEAR BOUNDARY CONDITIONS*

MINGXIN WANG[†‡] AND YONGHUI WU[†]

Abstract. This paper deals with the solutions of nonlinear parabolic equations $u_t = \nabla(a(u)\nabla u)$ with nonlinear boundary conditions $\partial u/\partial n = b(u)$ on $\partial\Omega\times(0,T)$, where $a(u)$ and $b(u)$ are positive and nondecreasing C^1 functions for $u > 0$. It is shown that if $\int^{+\infty} ds/b(s) < +\infty$ or $\int^{+\infty} ds/b(s) = +\infty$ and $\int^{+\infty} ds/(a(s)b(s)b'(s) + a(s)b(s) + a'(s)b^2(s)) < +\infty$, then the solution blows up in finite time, and the solution exists globally if $\int^{+\infty} ds/b(s) = +\infty$ and $\int^{+\infty} ds/(a(s)b(s)b'(s) + a(s)b(s) + a'(s)b^2(s)) = +\infty$.

Key words. nonlinear parabolic equations, nonlinear boundary conditions, global solutions, blowup, upper and lower solutions

AMS subject classifications. 35K55, 35K60, 35B40

1. Introduction.

The global existence or blow-up problems for parabolic equations are studied by many authors. Papers [1] and [2] studied the following equations:

$$\begin{aligned} &u_t = \Delta u, && x\in\Omega, \quad t>0, \\ &\frac{\partial u}{\partial n} = b(u), && x\in\partial\Omega, \quad t>0, \\ &u(x,0) = u_0(x) > 0, && x\in\bar\Omega, \end{aligned} \tag{1}$$

where n is the outer normal vector.

In [1] Levine and Payne proved that, for the special case $b(u) = |u|^{1+\epsilon}h(u)$, $\epsilon > 0$, $h(u)$ is increasing, problem (1) has no global solution. Later, Walter dealt with the generalized nonlinear boundary conditions and he obtained that if $b(u)$ and $b'(u)$ are continuous, positive, and increasing, then $\int^{+\infty} ds/b(s)b'(s) < +\infty$ implies that the solution of (1) blows up in finite time; $\int^{+\infty} ds/b(s)b'(s) = +\infty$ implies that the solution of (1) is bounded on $\bar\Omega \times [0,T]$ for any $T < +\infty$.

Recently, [3] considered the similar problem from another point of view, i.e., the nonlinear diffusion term.

$$\begin{aligned} &u_t = \nabla(a(u)\nabla u), && x\in\Omega, \quad t>0, \\ &\frac{\partial u}{\partial n} = 1, && x\in\partial\Omega, \quad t>0, \\ &u(x,0) = u_0(x) > 0, && x\in\bar\Omega. \end{aligned} \tag{2}$$

Let $a(u)$ and $a'(u)$ be continuous and positive; there exists $M > 0$ such that $\lim_{u\to+\infty}\sup a'(u)/a(u) \le M$. The author in [3] proved that if $\int^{+\infty} ds/a(s) < +\infty$,

* Received by the editors August 6, 1992; accepted for publication (in revised form) March 2, 1993.

[†] Institute of Systems Science, Academia Sinica, Beijing 100080, People's Republic of China.

[‡] The work of this author was supported by the National Natural Science Foundation of China.

then the solution of (2) blows up in finite time; if $\int^{+\infty} ds/a(s) = +\infty$, then the solution of (2) is bounded on $\bar{\Omega} \times [0,T]$ for any $T < +\infty$.

Because $\lim_{u\to+\infty} \sup a'(u)/a(u) \le M$, it is obvious that the convergence of $\int^{+\infty} ds/a(s)$ is equivalent to that of $\int^{+\infty} ds/(a(s)+a'(s))$.

In this paper, we combine the two nonlinear cases by considering the following nonlinear parabolic equations with nonlinear boundary conditions:

$$(3)\qquad \begin{aligned} &u_t = \nabla(a(u)\nabla u), && x\in\Omega,\quad t>0,\\ &\frac{\partial u}{\partial n} = b(u), && x\in\partial\Omega,\quad t>0,\\ &u(x,0) = u_0(x) > 0, && x\in\bar{\Omega}. \end{aligned}$$

Our approach depends heavily upon the upper and lower solution method which was first introduced by Walter in [2] to deal with nonlinear boundary problems. We obtain similar results of (1) and (2) and can deduce the results of (1) and (2) as our special cases.

2. Main results and an example. Throughout this section we suppose that $\Omega \subset \mathbb{R}^N$ is a smooth bounded domain, and the initial data $u_0(x)$ is a positive C^1 function. Moreover, we assume the following:

(H) $a(s)$ and $b(s)$ are positive and nondecreasing C^1 functions for $s > 0$.

Under the above assumptions, it is well known that (3) has a unique local solution $u(x,t)$ and it is positive on $\bar{\Omega} \times [0, T_{\max})$.

Before giving our main results, we first state a comparison theorem, which is a special case in [4, Thm. 31.IV].

PROPOSITION (comparison principle). Let u, v be two positive smooth functions satisfying

$$\begin{aligned} &u_t - \nabla(a(u)\nabla u) \le v_t - \nabla(a(v)\nabla v), && x\in\Omega,\quad t\in[0,T],\\ &\frac{\partial u}{\partial n} - b(u) \le \frac{\partial v}{\partial n} - b(v), && x\in\partial\Omega,\quad t\in[0,T],\\ &u(x,0)\le v(x,0), && x\in\bar{\Omega}. \end{aligned}$$

Then $u(x,t) \le v(x,t)$ for all $x \in \bar{\Omega}$, $t \in (0,T)$.

We state our main theorems and compare them with the previous results.

THEOREM 1. *Let* (H) *hold. If* $\int^{+\infty} ds/b(s) < +\infty$, *then the solution* $u(x,t)$ *of* (3) *blows up in finite time.*

THEOREM 2. *Let* (H) *hold,* $a(s) + a(s)b'(s) + a'(s)b(s)$ *be nondecreasing, and* $\int^{+\infty} ds/b(s) = +\infty$.

(i) If $\int^{+\infty} ds/(b(s)(a(s)+a(s)b'(s)+a'(s)b(s))) < +\infty$, then the solution $u(x,t)$ of (3) blows up in finite time.

(ii) If $\int^{+\infty} ds/(b(s)(a(s)+a(s)b'(s)+a'(s)b(s))) = +\infty$, then the solution $u(x,t)$ of (3) exists globally.

From Theorems 1 and 2, we see easily that if $a'(s)$ and $b'(s)$ are nondecreasing, then $a(s) + a(s)b'(s) + a'(s)b(s)$ is nondecreasing. Hence the solution $u(x,t)$ of (3) exists globally if and only if $\int^{+\infty} ds/b(s) = +\infty$ and $\int^{+\infty} ds/(b(s)(a(s) + a(s)b'(s) + a'(s)b(s))) = +\infty$.

COROLLARY 1. *Let $b(s) = 1$ and $a(s), a'(s)$ be continuous, positive, and nondecreasing. Then if $\int^{+\infty} ds/(a(s) + a'(s)) < +\infty$, the solution of* (3) *blows up in finite time; if $\int^{+\infty} ds/(a(s) + a'(s)) = +\infty$, the solution of* (3) *exists globally.*

This is the case of [3].

COROLLARY 2. *Let $a(s) = 1, b(s)$ and $b'(s)$ be continuous, positive, and nondecreasing. Then, $\int^{+\infty} ds/b(s)b'(s) < +\infty$ implies that the solution of* (3) *blows up in finite time; $\int^{+\infty} ds/b(s)b'(s) = +\infty$ implies that the solution of* (3) *exists globally.*

This is the case of [2].

Example. Let $a(u) = u^p, b(u) = u^q, p, q \geq 0$. If $q > 1$, the solution of (3) blows up in finite time. If $q \leq 1$, then $p + q > 1$ implies that the solution of (3) blows up in finite time; $p + q = 1$ implies that the solution of (3) exists globally; $p + q < 1$ implies $q < 1$; by a similar method to the proof of (ii) of Theorem 2, it can be proved that the solution of (3) exists globally.

3. Proofs of the main results. Now we are in the position to prove the results of §2. In this section we denote d as the diameter of the bounded domain $\Omega \subset \mathbb{R}^N$. Our steps are just the same as those of [2] and [3]. The basic idea is to construct a lower solution $\underline{u}(x,t)$ or an upper solution $\bar{u}(x,t)$ of (3).

Proof of Theorem 1. By the comparison principle, it is easy to prove that the solution $u(x,t)$ of (3) satisfies $u(x,t) \geq \epsilon_0 = \min_{\bar{\Omega}} u_0(x) > 0$. We first assume that there exists a constant $c_0 > 0$ such that $b'(u) \geq c_0$ for $u \geq \epsilon_0/4$. Let $\psi(s)$ be the solution of the following ordinary differential equation

$$\psi'(s) = b(\psi(s)), \qquad s > 0, \tag{4}$$

$$\psi(0) = \psi_0 = \tfrac{1}{4}\epsilon_0.$$

Then $\psi''(s) = b'(\psi(s))\psi'(s)$. Since $b(s)$ is continuous and $\int^{+\infty} ds/b(s) < +\infty$, it is easy to deduce that $\psi(s)$ exists locally and blows up in finite time. Because $b(u) > 0$ for $u > 0$, we find that there exists a unique $s_1 > 0$ such that

$$\psi(s_1) = \epsilon_0/2.$$

Now we construct a lower solution $\underline{u}(x,t)$ of (3) as follows.

$$\underline{u}(x,t) = \psi(\epsilon t + c_1(h^*(x) + A)),$$

$$h^*(x) = x_1 + \cdots + x_N, \qquad c_1 = \frac{1}{2}\min\left\{\frac{s_1}{2Nd+1}, \frac{1}{N}\right\},$$

$$\epsilon = \frac{1}{2}Na(\psi(0))c_0c_1^2, \qquad A = Nd + 1.$$

Then

$$\frac{\partial \underline{u}}{\partial n} = c_1\psi'(s)\frac{\partial h^*}{\partial n} = b(\psi(s))c_1\frac{\partial h^*}{\partial n} \leq \frac{1}{2}b(\underline{u}).$$

Since

$$|\nabla h^*|^2 = N \quad \text{and} \quad \Delta h^* = 0,$$

direct computations show that

$$\begin{aligned}
\underline{u}_t &= \epsilon\psi'(s),\\
\nabla(a(\underline{u})\nabla\underline{u}) &= c_1^2 a(\psi(s))\psi''(s)|\nabla h^*|^2 + c_1^2 a'(\psi(s))\psi'^2(s)|\nabla h^*|^2\\
&> c_1^2 a(\psi(s))\psi''(s)|\nabla h^*|^2\\
&= c_1^2 a(\psi(s))b'(\psi(s))\psi'(s)N.
\end{aligned}$$

Because $\psi(s)$ is increasing, we know that $\underline{u} = \psi(\epsilon t + c_1(h^*(x) + A)) > \psi(0) = \frac{1}{4}\epsilon_0$, $\underline{u}(x,0) = \psi(c_1(h^*(x) + A)) \le \psi(c_1(2Nd + 1)) \le \psi(s_1) = \frac{1}{2}\epsilon_0$. And hence

$$b'(\underline{u}) = b'(\psi(s)) \ge c_0,$$

so that

$$\nabla(a(\underline{u})\nabla\underline{u}) > c_1^2 c_0 N a(\psi(0))\psi'(s).$$

Therefore, we have that

$$\begin{aligned}
&\underline{u}_t \le \nabla(a(\underline{u})\nabla\underline{u}), && x \in \Omega, \quad t > 0,\\
&\frac{\partial \underline{u}}{\partial n} < b(\underline{u}), && x \in \partial\Omega, \quad t > 0,\\
&\underline{u}(x,0) < u_0(x), && x \in \bar{\Omega}.
\end{aligned}$$

This shows that $\underline{u}(x,t)$ is a lower solution of (3); therefore, $u(x,t) \ge \underline{u}(x,t)$. Since $\underline{u}(x,t) \ge \psi(\epsilon t), \psi(\epsilon t)$ blows up in finite time. So does $u(x,t)$.

If $b'(u)$ has no positive lower bound for $u \ge \epsilon_0/4$, we can construct a new function $b_1(u)$ having this property and satisfying the same conditions of $b(u)$ and $b_1(u) \le b(u)$.

In fact, let $\epsilon_1 = \frac{1}{8}\epsilon_0$. Since $b'(u) \ge 0, b(u) > 0$ and $\int^{+\infty} ds/b(s) \le c < +\infty$, we know that

$$\frac{1}{b(u)}(u - \epsilon_1) \le \int_{\epsilon_1}^{u} \frac{ds}{b(s)} \le c.$$

Hence $b(u) \ge (1/c)(u - \epsilon_1) \ge u/2c$ for $u \ge 2\epsilon_1 = \frac{1}{4}\epsilon_0$. Let

$$b_1(u) = \frac{1}{2\epsilon_1}\int_{u-\epsilon_1}^{u}\left(b(s) + \frac{1}{2c}s\right)ds,$$

then

$$\begin{aligned}
b_1(u) &\le \frac{1}{2}b(u) + \frac{1}{8c\epsilon_1}(2\epsilon_1 u - \epsilon_1^2)\\
&\le \frac{1}{2}\left(b(u) + \frac{1}{2c}u\right) \le b(u) \quad \text{for} \quad u \ge \epsilon_0/4,\\
b_1(u) &\ge \frac{1}{2\epsilon_1}\int_{u-\epsilon_1}^{u} b(s)ds \ge \frac{1}{2}b(u - \epsilon_1),\\
b_1'(u) &= \frac{1}{2\epsilon_1}(b(u) - b(u - \epsilon_1)) + \frac{1}{4c} \ge \frac{1}{4c}.
\end{aligned}$$

These facts show that $b_1(u)$ and $b_1'(u)$ are continuous, positive on $[\epsilon_0/4, +\infty)$ and $\int^{+\infty} ds/b_1(s) < +\infty$.

By the above proof, we know that the solution $u_1(x,t)$ of (3), corresponding to $b_1(u)$, blows up in finite time. Since $u_1(x,t)$ and $u(x,t)$ are both greater than or equal to $\epsilon_0(= 8\epsilon_1)$, we infer that $u(x,t) \geq u_1(x,t)$ from the comparison principle. And hence $u(x,t)$ blows up in finite time, which completes our proof.

Proof of Theorem 2. Let $h(x)$ be the solution of the following problem:

$$(5)\qquad \begin{aligned} &\Delta h(x) = k = \frac{|\partial\Omega|}{|\Omega|}, \qquad x \in \Omega,\\ &\frac{\partial h}{\partial n} = 1, \qquad x \in \partial\Omega. \end{aligned}$$

If $h(x) \in C^2(\bar{\Omega})$ is a solution of (5), it is easy to see that $h(x) + c$ is also a solution of (5) for any constant $c > 0$. Hence we assume that $h(x) > 0$ on $\bar{\Omega}$, and there exists a constant $L > 0$ such that

$$\max_{\bar{\Omega}}(h(x) + |\nabla h(x)|) \leq L.$$

(i) If $\int^{+\infty} ds/(b(s)(a(s) + a(s)b'(s) + a'(s)b(s))) < +\infty$. Let $\psi(s)$ be the solution of (4). At this time $\psi(s)$ exists for all s since $\int^{+\infty} ds/b(s) = +\infty$. It follows from $b(u) > 0$ and $b'(u) \geq 0$ that $\lim_{s\to+\infty} \psi(s) = +\infty$.

Let $g(t)$ be the solution of the ordinary differential equation

$$(6)\qquad \begin{aligned} &g'(t) = c_3[a(\psi(g)) + a(\psi(g))b'(\psi(g)) + a'(\psi(g))b(\psi(g))], \qquad t > 0,\\ &g(0) = g_0 = \tfrac{1}{3}s_1, \end{aligned}$$

where

$$c_2 = \min\left\{\frac{s_1}{3(2Nd+1)}, \frac{1}{4Nd}\right\},$$

$$\lambda = \min\left\{\frac{c_2}{2L}, \frac{1}{2}, \frac{s_1}{3L}\right\}, \qquad c_3 = \min\left\{k\lambda, \left(\frac{c_2}{2}\right)^2\right\},$$

$$s_1 > 0 \quad \text{such that} \quad \psi(s_1) = \tfrac{1}{2}\epsilon_0.$$

Since $\int^{+\infty} ds/(b(s)(a(s) + a(s)b'(s) + a'(s)b(s))) < +\infty$, we have that $g(t)$ exists locally and blows up in finite time.

Let

$$\underline{u}(x,t) = \psi(g(t) + \lambda h(x) + c_2(h^*(x) + A)),$$

where $h^*(x) = x_1 + \cdots + x_N, A = Nd + 1$. Denote

$$s = g(t) + \lambda h(x) + c_2(h^*(x) + A).$$

Then we have that

$$\underline{u}_t = \psi'(s)g'(t),$$

$$\begin{aligned}\nabla(a(\underline{u})\nabla\underline{u}) &= a(\underline{u})[\lambda\psi'(s)\Delta h+\psi''(s)|\lambda\nabla h+c_2\nabla h^*|^2]\\ &\quad +a'(\underline{u})\psi'^2(s)|\lambda\nabla h+c_2\nabla h^*|^2\\ &= k\lambda a(\underline{u})\psi'(s)+(a(\underline{u})b'(\underline{u})\psi'(s)+a'(\underline{u})\psi'^2(s))|\lambda\nabla h+c_2\nabla h^*|^2\\ &\geq \left[k\lambda a(\underline{u})+\left(\frac{c_2}{2}\right)^2(a(\underline{u})b'(\underline{u})+a'(\underline{u})b(\underline{u}))\right]\psi'(s)\\ &\geq c_3[a(\psi(g))+a(\psi(g))b'(\psi(g))+a'(\psi(g))b(\psi(g))]\psi'(s)\end{aligned}$$

since $\psi(s)>\psi(g)$ and $a(s)+a(s)b'(s)+a'(s)b(s)$ is nondecreasing.

Because $\psi'(s)=b(\psi(s))>0$, it follows from (6) that

$$\underline{u}_t\leq\nabla(a(\underline{u})\nabla\underline{u}).$$

Moreover,

$$\frac{\partial\underline{u}}{\partial n}=\psi'(s)\left(\lambda\frac{\partial h}{\partial n}+c_2\frac{\partial h^*}{\partial n}\right)\leq\psi'(s)(\lambda+c_2Nd)<b(\underline{u}),$$

$$\begin{aligned}\underline{u}(x,0) &= \psi(g_0+\lambda h(x)+c_2(h^*(x)+A))\\ &< \psi\left(\frac{1}{3}s_1+\lambda L+c_2(2Nd+1)\right)\\ &< \psi(s_1)=\frac{1}{2}\epsilon_0<u_0(x).\end{aligned}$$

Therefore, by the comparison principle we obtain that

$$u(x,t)\geq\underline{u}(x,t).$$

Since $g(t)$ blows up in finite time and $\lim_{s\to+\infty}\psi(s)=+\infty$, we get that $\underline{u}(x,t)$ blows up in finite time. So does $u(x,t)$.

(ii) If $\int^{+\infty}ds/(b(s)(a(s)+a(s)b'(s)+a'(s)b(s)))=+\infty$. Let $\psi(s)$ be the solution of the following ordinary differential equation:

$$\begin{aligned}&\psi'(s)=b(\psi(s)), \qquad s>0,\\ &\psi(0)=\psi_0=\max_{\Omega}u_0(x).\end{aligned}\tag{7}$$

Then $\psi(s)$ is global solution of (7) because of $\int^{+\infty}ds/b(s)=+\infty$.

Let $g(t)$ be the solution of the following problem:

$$\begin{aligned}g'(t) &= (k+L^2)(a(\psi(g+L))+a(\psi(g+L))b'(\psi(g+L))\\ &\qquad +a'(\psi(g+L))b(\psi(g+L))), \qquad t>0,\\ g(0) &= 1.\end{aligned}$$

$\int^{+\infty} ds/(b(s)(a(s) + a(s)b'(s) + a'(s)b(s))) = +\infty$ implies that $g(t)$ exists globally. We construct $\bar{u}(x,t)$ as follows:

$$\overline{u}(x,t) = \psi(g(t) + h(x)).$$

Through direct computations, it is easy to see that

$$\begin{aligned} &\bar{u}_t \geq \nabla(a(\overline{u})\nabla\bar{u}), && x \in \Omega, \quad t > 0, \\ &\frac{\partial\bar{u}}{\partial n} > b(\bar{u}), && x \in \partial\Omega, \quad t > 0, \\ &\bar{u}(x,0) > u_0(x), && x \in \bar{\Omega}. \end{aligned}$$

This shows that $\bar{u}(x,t)$ is an upper solution of (3). Since ψ and g exist globally, we know that $\bar{u}(x,t)$ exists for all $t > 0$. So does $u(x,t)$. This concludes the proof of Theorem 2.

REFERENCES

[1] H. A. Levine and L. E. Payne, *Nonexistence theorems for the heat equation with nonlinear boundary condition and for the porous medium equation backward in time*, J. Differential Equations, 16 (1974), pp. 319–334.

[2] W. Walter, *On existence and nonexistence in the large of solutions of parabolic differential equation with a nonlinear boundary condition*, SIAM J. Math. Anal., 6 (1975), pp. 85–90.

[3] Yin Hong-ming, *Blow-up versus global solvability for a class of nonlinear parabolic equations*, preprint.

[4] W. Walter, *Differential and Integral Inequalities*, Ergebnisse der Mathematik Und ihrer Grenzgebiete, Band 55, Springer-Verlag, New York, Heidelberg, Berlin, 1970.

SIAM J. MATH. ANAL.
Vol. 24, No. 6, pp. 1522–1536, November 1993

HIGHER-ORDER REGULARITY FOR THE SOLUTIONS OF SOME DEGENERATE QUASILINEAR ELLIPTIC EQUATIONS IN THE PLANE*

W. B. LIU[†] AND JOHN W. BARRETT[†]

Abstract. Local $C^{k,\beta}$ and $W^{2+k,v}$ ($k \geq 1, \beta > 0$, and $v \geq 1$) regularity is established for the solutions of a class of degenerate quasilinear elliptic equations, which include the p-Laplacian. Unlike the known local regularity results for such equations, k is larger than 2 in many notable cases. These results generalize those in [13], which were established only for the p-Laplacian. Furthermore, local results are extended to obtain a global regularity result in some cases. Global results of this type are essential in proving optimal error bounds for the finite element approximation of such equations.

Key words. regularity, degenerate, elliptic equations, quasilinear

AMS subject classifications. 35B65, 35J70, 35J60

1. Introduction.

Let Ω be a bounded open set in R^2 with a Lipschitz boundary $\partial\Omega$. Let φ be a positive continuous function on $(0, \infty)$. Many mathematical models from physical processes have the following form: given ρ, find u such that

$$-\nabla \cdot (\varphi(|\nabla u|)\nabla u) = \rho \quad \text{in } \Omega \tag{1.1}$$

plus some boundary conditions. Such models arise in fluid mechanics (see [4] and [5]), nonlinear diffusion (see [24]), and nonlinear elasticity (see [3]). An example of the latter is the mode III problem, longitudinal shear, for a power law material (see [3]). This leads to (1.1) with $\varphi(t) = t^{p-2}$ ($1 < p \leq 2$), which is usually referred to as the p-Laplacian.

There is much work on the regularity of the solutions of (1.1) if $\varphi > 0$ is smooth on $[0, \infty)$ and satisfies certain coercivity and growth conditions. In this case the solutions of (1.1) will generally be smooth in any open set $G \subset\subset \Omega$. As key references, one can refer to [11], [12], [15], and [23]. These results are generally true for (1.1) when $\Omega \subset R^n$.

However, as seen above, in some physically relevant models φ may not be smooth on $[0, \infty)$ or may not satisfy the coercivity and growth conditions. Equations of this type are usually referred to as degenerate. A typical example is the p-Laplacian with $p \neq 2$. The well established regularity results for nondegenerate equations are generally not applicable to such cases (see [29] for some counterexamples). The regularity theory for these degenerate equations is much more complicated and until quite recently very little was known (see [31]).

Since the publication of [30], extensive work has been done on the regularity of the solutions of (1.1) and more general degenerate quasilinear equations and systems. One can find such work, for example, in [1], [2], [9], [10], [13], [17]–[19], [20]–[22], and [26]–[29]. Generally speaking, local and global $C^{1,\alpha}$ regularity has been established for the solutions of (1.1) and more general systems (for example, see [9], [18], [19], and [29]). In addition, some local H^2 regularity results have been established; for example, see [26] and [29]. These results remain true for (1.1) when $\Omega \subset R^n$. It should be noted that it is usually quite difficult to extend a local regularity result to a

* Received by the editors May 18, 1992; accepted for publication (in revised form) March 1, 1993.

[†] Department of Mathematics, Imperial College, London SW7 2BZ, United Kingdom. The work of the first author was supported by Science and Engineering Research Council of the United Kingdom grant GR/F81225.

global one due to the degenerate nonlinearity. In the plane there exist more powerful methods to study (1.1); for example, complex function theory. In the case $\varphi(t) = t^{p-2}$ and $\rho = 0$ a local $W^{2,2+\eta}(\Omega)$ regularity result, where $\eta(p) > 0$ has been proved, using the theory of quasi-regular mappings, firstly for $p \geq 2$ in [7] and then for $p > 1$ in [22]. In [20] global H^2 regularity results have been established for (1.1), for $\rho \in L^r(\Omega)$, $r > 2$, when $p \in (1, 2]$ and for $\rho = 0$ when $p > 2$, by using an a priori estimate for the solutions of a linear elliptic equation in the plane with only $L^\infty(\Omega)$ coefficients.

These global H^2 regularity results have been used in the error analysis of the finite element approximation of (1.1) in [6], [20], and [21]. From this work it is seen that the regularity of the solutions has a considerable influence on the order of convergence that can be proved for their finite element approximation. Taking the p-Laplacian $(1 < p < 2)$ as an example, if $u \in W^{2,p}(\Omega)$ one can prove that the error for the continuous piecewise linear finite element approximation converges at the rate $h^{p/2}$ in $W^{1,p}(\Omega)$, whereas one can prove that it converges at the optimal rate h if $u \in C^{2,2/p-1}(\bar{\Omega}) \cap W^{3,1}(\Omega)$, see [6], or $u \in W^{3,p}(\Omega)$. With some restrictions on ρ and the boundary data, the global regularity $u \in W^{1+2/p,p}(\Omega)$ is shown in [21] to be sufficient to ensure the optimal convergence rate h. Thus one can see that it is important to establish some global higher-order regularity results in order to obtain optimal error bounds in $W^{1,p}(\Omega)$ for the continuous piecewise linear finite element approximation of (1.1).

A first step in this direction was taken in [13], where $C^{k,\alpha}$ and $W^{2+k,q}$ regularity has been established, using the theory of quasi-regular mappings for the solutions of the p-Laplacian in the plane with $\rho = 0$, although the results are only local. Unlike most known results, k is larger than 2 in many interesting cases; for example, the p-Laplacian with $1 < p \leq 2$. From this result it was pointed out in [21] that global $W^{3,p}(\Omega)$, $C^{2,2/p-1}(\bar{\Omega})$, and $W^{1+2/p,p}(\Omega)$ regularity can indeed be achieved for solutions of the p-Laplacian in some physically relevant cases, though this result is only applicable to a limited class of boundary data. It should be noted that the results in [13] are sharp and such high-order regularity (local or global) is not generally true for the case $\rho \neq 0$; see, for example, [17].

The purpose of this paper is to generalize the results in [13] to a class of degenerate quasilinear elliptic equations, which includes the p-Laplacian. Furthermore, we show that this local result can be extended to a global one in some cases.

Throughout this paper, we adopt the standard notation $W^{s,v}(\Omega)(s \geq 0,\ v \in [1, \infty])$ for Sobolev spaces on Ω with norm $\|\cdot\|_{W^{s,v}(\Omega)}$; see [14], for example. We denote $W^{m,2}(\Omega)$ by $H^m(\Omega)$ for nonnegative integers m and the spaces $W^{s,v}_{\mathrm{loc}}(\Omega)$ consist of all functions which are in $W^{s,v}(D)$ for all open sets $D \subset\subset \Omega$ (that is, $\bar{D} \subset \Omega$). For the definitions of the spaces $C^{k,\beta}(\Omega)$ (k a nonnegative integer and $0 \leq \beta \leq 1$) and domains with $C^{k,\beta}$ boundaries; see [12], for example.

2. Preliminaries.

Let Ω be a bounded open set in R^2 with a Lipschitz boundary $\partial\Omega$. Let φ satisfy the following conditions:

(A1) The function $t \rightarrow \varphi(t)t$ belongs to $C^0[0, \infty) \cap C^\infty(0, \infty)$.

(A2) There exist constants $q > 1$, $q \neq 2$, and $C_i, \kappa > 0$ such that for all $t \geq 0$

$$C_1 t \min\{t^{q-2}, (1+t)^{q-2}\} \leq \varphi(t)t \leq C_2 t \max\{t^{q-2}, (1+t)^{q-2}\}, \tag{2.1a}$$

and $\Gamma(t) \equiv \varphi'(t)t/\varphi(t)$ is such that for all $t \geq 0$,

$$\kappa \leq \Gamma(t)/(q-2) \leq 1. \tag{2.1b}$$

(A3) The function $\Gamma(t)$ is analytic at $t = 0$; that is, in an interval about $t = 0$,

$$\Gamma(t) \equiv \sum_{i=0}^{\infty} \gamma_i t^i. \tag{2.2a}$$

In addition we assume that there is a $\eta > 0$ such that

$$|\gamma_i| \leq |\gamma_0|\eta^i \quad \text{for all } i \geq 1, \tag{2.2b}$$

where we note from (2.1a) that $\Gamma(0) \equiv \gamma_0 \neq 0$.

We note that the regularity requirement $C^{\infty}(0, \infty)$ in (A1) can be appropriately relaxed and the results in this paper remain valid.

In this paper we will study the local and global regularity of the weak solutions $u \in W^{1,q}_{\text{loc}}(\Omega)(q > 1)$ of the following equation:

$$\nabla \,.\, (\varphi(|\nabla u|)\nabla u) = 0 \quad \text{in } \Omega. \tag{2.3}$$

Example 2.1. For example, it is easy to see that $\varphi(t) \equiv t^{q-2}$ or $\varphi(t) \equiv [t(1 + t)]^{(q-2)/2}$ or $\varphi(t) \equiv [t(1 + t^2)]^{(q-2)/3}(q > 1)$ satisfies the conditions (A1)–(A3). It is also easy to show that if φ satisfies the conditions (A1) and (A2), then it satisfies the structural conditions (1.2)–(1.5) in [27] and (1.3)–(1.5) in [29].

Remark 2.1. We note that if $\varphi(t)$ satisfies the conditions (A1)–(A3) and u satisfies (2.3), then for any constant $\lambda > 0$, $\equiv \lambda^{-1}u$ is such that $\nabla \,.\, (\hat{\varphi}(|\nabla \hat{u}|)\nabla \hat{u}) = 0$ in Ω with $\hat{\varphi}(t) \equiv \varphi(\lambda t)$. Therefore $\hat{\varphi}(t)$ satisfies (A1)–(A3) with $\hat{\Gamma}(t) \equiv \hat{\varphi}'(t)t/\hat{\varphi}(t) \equiv \Gamma(\lambda t)$, $\hat{\Gamma}(t) \equiv \sum_{i=0}^{\infty} \hat{\gamma}_i t^i$, and $|\hat{\gamma}_i| \leq |\hat{\gamma}_0|(\lambda\eta)^i$ for all $i \geq 1$. Therefore without loss of generality we can assume that $\eta > 0$ in (2.2b) is as small as we please throughout this paper.

In [8], [19], [27], and [29] the existence of solutions, $u \in W^{1,q}_{\text{loc}}(\Omega)$, to the weak formulation of (2.3) is established. It follows from [9] and [29] that $u \in C^{1,\alpha}(\Omega)$ for an $\alpha > 0$. Furthermore, if some suitable boundary conditions are imposed, the weak solution $u \in W^{1,q}(\Omega)$ is unique.

Let u be a weak solution of (2.3), and let $f \equiv u_x - iu_y$ be the complex gradient of u. It turns out that regularity of f, and hence u, can be more easily examined if we view it as a function of a complex variable. To this end we first express (2.3) in the form of a complex equation. Let us recall for $z \equiv x + iy$, $(x, y) \in R^2$ that $\partial/\partial z \equiv (\partial/\partial x - i\partial/\partial y)/2$ and $\partial/\partial \bar{z} \equiv (\partial/\partial x + i\partial/\partial y)/2$.

LEMMA 2.1. *The mapping $f \equiv u_x - iu_y$ is K-quasi-regular with $K = \max\{q - 1, 1/(q - 1)\}$. Hence $f \in [W^{1,2+\eta}_{\text{loc}}(\Omega)]^2$ for a $\eta > 0$ and $f^{-1}(0) \equiv \{z \in \Omega : \nabla u(z) = 0\}$ is discrete provided that u is not identically constant.*

Let $p \equiv 2 + \Gamma(0)$; then

$$p \in [q, 2) \quad \text{if } q \in (1, 2) \quad \text{and} \quad p \in (2, q] \quad \text{if } q > 2. \tag{2.4}$$

On setting $G(t) \equiv \Gamma(t)/\Gamma(0)$ and $F(t) \equiv pG(t)/(2 + (p - 2)G(t))$, it follows that f satisfies for almost every $z \in \Omega$,

$$\partial f/\partial \bar{z} = (1/p - \tfrac{1}{2})[(\bar{f}/f)(\partial f/\partial z) + (f/\bar{f})(\overline{\partial f/\partial z})]F(|f|). \tag{2.5}$$

Proof. We will show that f is a quasi-regular mapping by generalizing the approach given in [22] for the p-Laplacian. Firstly we prove that $u \in H^2_{\text{loc}}(\Omega)$.

Let $\varphi^\varepsilon(t) \equiv \varphi((t^2+\varepsilon)^{1/2})$ and u^ε be the unique solution of the following problem:

$$(2.6) \qquad \nabla \cdot (\varphi^\varepsilon(|\nabla u^\varepsilon|)\nabla u^\varepsilon) = 0 \quad \text{in } B \quad \text{and} \quad u^\varepsilon = u \quad \text{on } \partial B,$$

where B is an open ball in Ω. As $u \in C^1(\bar{B})$, it follows from [12] and [15] that $u^\varepsilon \in H^2_{\text{loc}}(B) \cap C^1(B)$ (actually $u^\varepsilon \in C^{2,\beta}(B)$ for a $\beta > 0$). Noting that our Assumptions (A1) and (A2) are equivalent to (1.2)–(1.5) in [27], it follows that $u_\varepsilon \to u$ in $W^{1,q}(B)$ as $\varepsilon \to 0$. On the other hand, (2.6) can be restated as

$$(2.7)\ a^\epsilon\partial^2 u^\epsilon/\partial x^2 + 2b^\epsilon\partial^2 u^\epsilon/\partial x\partial y + c^\epsilon\partial^2 u^\epsilon/\partial y^2 = 0 \quad \text{in } B \quad \text{and} \quad u^\epsilon = u \quad \text{on } \partial B,$$

where $v \equiv (\epsilon + |\nabla u^\epsilon|^2)^{1/2}, v_1 \equiv \partial u^\epsilon/\partial x, v_2 \equiv \partial u^\epsilon/\partial y, a^\epsilon \equiv [1 + [\varphi'(v)v/\varphi(v)](v_1/v)^2]$, $b^\epsilon \equiv [\varphi'(v)v/\varphi(v)](v_1v_2/v^2)$, and $c^\epsilon \equiv [1 + [\varphi'(v)v/\varphi(v)](v_2/v)^2]$. Setting $d^\epsilon \equiv a^\epsilon c^\epsilon - (b^\epsilon)^2 \equiv 1 + [\varphi'(v)v/\varphi(v)][\{(v_1)^2 + (v_2)^2\}/v^2]$, we have from (2.1b) that

$$(2.8) \qquad \min\{1, q-1\} \le a^\epsilon, c^\epsilon, d^\epsilon \le \max\{1, q-1\} \quad \text{and} \quad |b^\epsilon| \le |q-2|;$$

that is, the equation is uniformly (with respect to ϵ) elliptic with bounded coefficients.

We now apply the inequality (17.20) in [15, p. 228]. For any $B'' \subset\subset B' \subset\subset B$ this yields that $|u^\epsilon|_{H^2(B'')} \le C\|u^\epsilon\|_{H^1(B')}$, where C is independent of ϵ. In [27] it has been shown that $|u^\epsilon| + |\nabla u^\epsilon|$ is bounded on B' as $\epsilon \to 0$ since $u \in C^0(\bar{B})$. This implies that $\|u^\epsilon\|_{H^2(B'')}$ is uniformly bounded. Hence $u \in H^2(B'')$ so that $u \in H^2_{\text{loc}}(\Omega)$, since u^ϵ will converge to u weakly in $H^2(B'')$ as $\epsilon \to 0$. Since the proof of the uniform boundedness of $|u^\epsilon| + |\nabla u^\epsilon|$ in [27] is quite complicated, we now give a simpler proof. Let $u_0 = \min_{x\in\partial B} u(x)$ and $u_1 = \max_{x\in\partial B} u(x)$. Note that u_0 and u_1 are constants so that one can apply Lemma 3 in [27] to (2.6). This shows that $u_0 \le u^\epsilon \le u_1$ for any $\epsilon > 0$. Then by applying Theorem 12.4 in [12] with $f = 0$ and the interpolation inequality (12.23)[(6.82)] in [12] one can see that $|u^\epsilon| + |\nabla u^\epsilon|$ is bounded on B' as $\epsilon \to 0$. Of course for $q > 2$, $\|u^\epsilon\|_{H^1(B)}$ is uniformly bounded since $\|u^\epsilon\|_{W^{1,q}(B)}$ is. Therefore, we have that $u \in H^2_{\text{loc}}(\Omega)$, which implies that $f \in [H^1_{\text{loc}}(\Omega)]^2$.

Let $f^\epsilon \equiv (v_1, -v_2)$. Adopting the same notation as in Theorem 1 in [22] we have from (2.8) that for all $z \in B$,

$$
\begin{aligned}
(2.9)\quad & \\
&\|Df^\epsilon(z)\|^2/Jf^\epsilon(z)\\
&\quad \equiv [(\partial v_1/\partial x)^2 + (\partial v_1/\partial y)^2 + 2(\partial v_1/y)^2](z)/[(\partial v_1/\partial y)^2 - (\partial v_1/\partial x)(\partial v_2/\partial y)](z)\\
&\quad \le [(a^\epsilon)^2 + 2(b^\epsilon)^2 + (c^\epsilon)^2](z)/[a^\epsilon c^\epsilon - (b^\epsilon)^2](z)\\
&\quad = [1 + (d^\epsilon)^2](z)/d^\epsilon(z) \le 2K,
\end{aligned}
$$

where $K = \max\{q-1, 1/(q-1)\}$. Hence f^ϵ is a K-quasi-regular mapping; see [13] and [22]. As $|u^\epsilon| + |\nabla u^\epsilon|$ is bounded on B' as $\epsilon \to 0$, we can apply the argument in Theorem 1 in [22] and let $\epsilon \to 0$ in the above to yield that f is K-quasi-regular with $K = \max\{q-1, 1/(q-1)\}$.

It follows immediately from the theory of quasi-regular maps, see [15], that $f \in [W^{1,2+\eta}_{\text{loc}}(\Omega)]^2$ for a $\eta > 0$ and $f^{-1}(0) \equiv \{z \in \Omega : \nabla u(z) = 0\}$ is discrete, provided that u is not identically constant.

The result (2.4) follows immediately from Assumption (A2), which in turn implies that G and F are well defined. Finally we note that if $|\nabla u(z)| > 0$, then (2.3) yields that

(2.10)
$$2\partial f(z)/\partial \bar{z} \equiv \Delta u(z) = -\Gamma(|\nabla u(z)|)[(u_x)^2 u_{xx} + 2u_x u_y u_{xy} + (u_y)^2 u_{yy}](z)/|\nabla u(z)|^2$$
$$= (2-p)G(|\nabla u(z)|)[(u_x)^2 u_{xx} + 2u_x u_y u_{xy} + (u_y)^2 u_{yy}](z)/|\nabla u(z)|^2.$$

In addition, if $|\nabla u(z)| > 0$, then $2\partial f/\partial z \equiv u_{xx} - u_{yy} - 2iu_{xy}$, and so

$$(2.11)\qquad \begin{aligned}&[(\bar{f}/f)(\partial f/\partial z) + (f/\bar{f})(\overline{\partial f/\partial z})](z)\\ &\equiv -\Delta u(z) + 2[(u_x)^2 u_{xx} + 2u_x u_y u_{xy} + (u_y)^2 u_{yy}](z)/|\nabla u(z)|^2.\end{aligned}$$

Combining (2.10), (2.11), and noting that $\{z \in \Omega : \nabla u(z) = 0\}$ is discrete yields the desired result (2.5). □

We note that in the case $\varphi(t) = t^{q-2}, q \in (1, \infty)$, it was first proved in [7] for $q > 2$ and then in [22] for $q > 1$ that the corresponding map f was quasi regular.

It follows from classical regularity theory that the solution u of (2.3), and the corresponding f, is C^∞ outside the set $\{z \in \Omega : \nabla u(z) = 0\}$ (see, for example, [12]). Therefore it is sufficiently general to examine the regularity of u (or f) near one of the zero points of f, which one may assume is the origin. From now on, we will assume that $0 \in \Omega$ and $f(0) = 0$.

We now consider the regularity of local solutions of (2.5). We generalize the approach for the p-Laplacian given in [13]. From the theory of quasi-regular mappings, see [16], and as $f(0) = 0$ with $0 \in \Omega$ is an isolated zero of f, there exists a quasi-conformal homeomorphism χ defined in a neighborhood of $z = 0$ and a positive integer n such that $\chi(0) = 0$ and

$$(2.12)\qquad f(z) \equiv [\chi(z)]^n.$$

From (2.5) and (2.12) we have that

$$\partial\chi/\partial\bar{z} = (1/p - \tfrac{1}{2})[(\bar{\chi}^n/\chi^n)(\partial\chi/\partial z) + (\chi/\bar{\chi})(\overline{\partial\chi/\partial z})]F(|\chi|^n).$$

Let $H(\xi)$ be the inverse of $\chi(z)$ in a neighborhood of $z = 0$. It follows that H is quasi-conformal in a neighborhood B of $\xi = 0$, and hence $H \in C(B) \cap H^1(B)$ and $H \in C^\infty(B - \{0\})$. Then we have that

$$(2.13)\qquad \partial H/\partial\bar{\xi} = (\tfrac{1}{2} - 1/p)[(\bar{\xi}^n/\xi^n)(\overline{\partial H/\partial\xi}) + (\xi/\bar{\xi})(\partial H/\partial\xi)]F(|\xi|^n).$$

Let $\xi = re^{i\theta}$ and $H(r, \theta) \equiv H(re^{i\theta})$. Proceeding as in [13] one obtains that

$$(2.14)\qquad 2rH_r = -i(2 + (p-2)G(r^n))H_\theta + i(p-2)G(r^n)e^{-2in\theta}\overline{H_\theta}.$$

Expanding $H(r, \theta)$ into Fourier series with respect to θ, $0 \le \theta < 2\pi$,

$$(2.15)\qquad H(r, \theta) \equiv \sum_{k=-\infty}^{\infty} h_k(r)e^{i(k-n)\theta},$$

where $h_k \in C[0, \delta] \cap C^\infty(0, \delta]$ for $\delta > 0$. It follows from (2.14) and (2.15) that for all integers k,

$$(2.16a)\qquad 2rh_k'(r) = -(n-k)(2 + (p-2)G(r^n))h_k(r) - (n+k)(p-2)G(r^n)\overline{h_{-k}(r)}$$

and

$$(2.16\text{b}) \quad 2r\overline{h'_{-k}(r)} = -(n+k)(2+(p-2)G(r^n))\overline{h_{-k}(r)} - (n-k)(p-2)G(r^n)h_k(r).$$

Setting $X_k(r) \equiv (h_k(r), \overline{h_{-k}(r)})^T$, this system of equations can be restated as follows: Let n be a positive integer and $p \in (1,2) \cup (2,\infty)$. Then for any integer k find $X_k(r)$ such that

$$(2.17\text{a}) \qquad 2rX'_k(r) = -[A(k) + \hat{g}(r)B(k)]X_k(r),$$

where

(2.17b)

$$A(k) \equiv \begin{pmatrix} (n-k)p & (n+k)(p-2) \\ (n-k)(p-2) & (n+k)p \end{pmatrix}, \qquad B(k) \equiv \begin{pmatrix} (n-k) & (n+k) \\ (n-k) & (n+k) \end{pmatrix},$$

and

$$(2.17\text{c}) \qquad \hat{g}(r) \equiv (p-2)(G(r^n)-1).$$

The advantage of this reformulation is that the structure of the general solutions for the system $2rW'_k(r) = -A(k)W_k(r)$ is easily established when $k \neq -n$ and the function $\hat{g}(r)B(k)/r$ is regular.

Therefore, to study the regularity of local solutions of (2.5) we need to study the structure of the general solutions about $r = 0$ to the system (2.17). This we do in the next section.

3. Series expansion for $H(\xi)$. We will analyze the system (2.17) by developing a series solution about $r = 0$. In fact, we analyze a more general system: (2.17a, b) with $\hat{g}(r)$ being any real analytical function at $r = 0$ satisfying $\hat{g}(0) = 0$, instead of the specific choice (2.17c). First, we need a lemma.

LEMMA 3.1. *Let m be a positive integer and let $M \in R$ be such that $M \geq 1 + 2^m$. Then for any integer $k \in [0, m]$ and any integer $i \geq 1$,*

$$(3.1) \qquad \sum_{j=0}^{i-1} (i-j)^k M^j \leq M^i.$$

Proof. Simple calculations yield that (3.1) is true for (i) $k = 0$ for any integer $i \geq 1$, and (ii) $i = 1$ for any integer $k \in [0, m]$. Now assume (3.1) holds for all $k \in [0, \hat{m}]$ and any integer $i \geq 1$ with $0 \leq \hat{m} < m$. Suppose that for $k = \hat{m}+1$ it is also true for any integer $i \in [1, \hat{\imath}]$ with $\hat{\imath} \geq 1$. Then we have that

$$\begin{aligned} \sum_{j=0}^{\hat{\imath}} (\hat{\imath}+1-j)^{\hat{m}+1} M^j &= (\hat{m}+1)! \sum_{j=0}^{\hat{\imath}} M^j \sum_{k=0}^{\hat{m}+1} (\hat{\imath}-j)^k \Big/ [k!(\hat{m}+1-k)!] \\ &= M^{\hat{\imath}} + \sum_{k=0}^{\hat{m}+1} \{(\hat{m}+1)!/[k!(\hat{m}+1-k)!]\} \sum_{j=0}^{\hat{\imath}-1} (\hat{\imath}-j)^k M^j \\ &\leq \left[1 + \sum_{k=0}^{\hat{m}+1} \{(\hat{m}+1)!/[k!(\hat{m}+1-k)!]\}\right] M^{\hat{\imath}} \\ &\leq (1+2^{\hat{m}+1})M^{\hat{\imath}} \leq (1+2^m)M^{\hat{\imath}} \leq M^{\hat{\imath}+1}. \end{aligned}$$

Therefore, (3.1) also holds for $k = \hat{m}+1$ and $i = \hat{i}+1$. Hence by induction (3.1) holds for any integer $k \in [0, m]$ and any integer $i \geq 1$. □

THEOREM 3.1. *Let $p \in (1,2) \cup (2,\infty)$ and n be a positive integer. Suppose that $\hat{g}(r) \equiv \sum_{i=1}^{\infty} g_i r^i$ near $r = 0$ and that there exist $\mathcal{G} \in R^+$ and a positive integer m such that $|g_i| \leq \mathcal{G} i^m$ for all $i \geq 1$. Then for any integer k, except $k = -n$, the general solution $X_k(r)$ of* (2.17 a, b) *near $r = 0$ is of the form*

$$X_k(r) \equiv C_k^+ r^{\lambda_k^+} X_k^+(r) + C_k^- r^{\lambda_k^-}[X_k^-(r) + cr^{\rho_k}\ln(r)X_k^+(r)], \tag{3.2a}$$

where $C_k^{\pm}$ are general constants,

$$\lambda_k^{\pm} \equiv \{-pn \pm [4k^2(p-1) + n^2(p-2)^2]^{1/2}\}/2, \tag{3.2b}$$

$$\rho_k \equiv (\lambda_k^+ - \lambda_k^-) \equiv [4k^2(p-1) + n^2(p-2)^2]^{1/2}, \tag{3.2c}$$

$$\epsilon_k^{\pm} \equiv [(n-k)p + 2\lambda_k^{\pm}]/[(n+k)(2-p)]. \tag{3.2d}$$

The constant c is zero if ρ_k is not an integer, but possibly nonzero otherwise,

$$X_k^{\pm}(r) \equiv \begin{pmatrix} \hat{a}_k^{\pm}(r) + \hat{b}_k^{\pm}(r) \\ \epsilon_k^+ \hat{a}_k^{\pm}(r) + \epsilon_k^- \hat{b}_k^{\pm}(r) \end{pmatrix}, \tag{3.2e}$$

where $\hat{a}_k^{\pm}(0) + \hat{b}_k^{\pm}(0) = 1$ and

$$\hat{a}_k^{\pm}(r) \equiv \sum_{i=0}^{\infty} a_i^{\pm}(k) r^i, \qquad \hat{b}_k^{\pm}(r) \equiv \sum_{i=0}^{\infty} b_i^{\pm}(k) r^i. \tag{3.2f}$$

In addition there exist constants $C(k)$ and $M \geq 1 + 2^m$, dependent on p and n such that $|a_i^{\pm}(k)|, |b_i^{\pm}(k)| \leq C(k) M^i$ for all $i \geq 0$.

Proof. Let $\mathcal{X}(r)$ be a fundamental solution to the system $2rW_k'(r) = -A(k)W_k(r)$ when $k \neq n$; that is,

$$\mathcal{X}_k(r) \equiv \begin{pmatrix} r^{\lambda_k^+} & r^{\lambda_k^-} \\ \epsilon_k^+ r^{\lambda_k^+} & \epsilon_k^- r^{\lambda_k^-} \end{pmatrix},$$

where $\lambda_k^{\pm}$ are the eigenvalues of $-A(k)/2$ and $(1, \epsilon_k^{\pm})^T$ the corresponding eigenvectors; see (3.2b, d). We now seek two linearly independent solutions of (2.17a, b) of the form $\mathcal{X}_k(r)Y_k(r)$. Inserting $\mathcal{X}_k(r)Y_k(r)$ into (2.17a) one obtains

$$2rY_k'(r) = -\hat{g}(r)\mathcal{X}_k^{-1}(r)B(k)\mathcal{X}_k(r)Y_k(r)$$

and hence

$$Y_k'(r) = [\hat{g}(r)/r]D_k(r)Y_k(r), \tag{3.3}$$

where

$$D_k(r) \equiv \begin{pmatrix} d_1(k) & d_2(k)r^{-\rho_k} \\ d_3(k)r^{\rho_k} & d_4(k) \end{pmatrix}$$

$$\equiv (\epsilon_k^- - \epsilon_k^+)^{-1} \begin{pmatrix} (\epsilon_k^- - 1)\tau_k^+ & (\epsilon_k^- - 1)\tau_k^- r^{-\rho_k} \\ (1-\epsilon_k^+)\tau_k^+ r^{\rho_k} & (1-\epsilon_k^+)\tau_k^- \end{pmatrix}$$

with

$$\tau_k^\pm \equiv (n-k) + \epsilon_k^\pm(n+k).$$

It follows that there exists a $\mathcal{D} > 1$, independent of p, n, and k, such that

$$|d_j(k)| \le \mathcal{D}(|k|+n), \qquad j = 1 \to 4. \tag{3.4}$$

From our assumptions we have that $\hat{g}(r)/r \equiv \sum_{i=0}^{\infty} g_{i+1}r^i$. We now look for two linearly independent solutions $Y_k^\pm$ to (3.3) of the form

$$Y_k^\pm(r) \equiv \begin{pmatrix} r^{\alpha_k^\pm}\, \hat{a}_k^\pm(r) \\ r^{\beta_k^\pm}\, \hat{b}_k^\pm(r) \end{pmatrix} \equiv \begin{pmatrix} \sum\limits_{i=0}^{\infty} a_i^\pm(k) r^{\alpha_k^\pm + i} \\ \sum\limits_{i=0}^{\infty} b_i^\pm(k) r^{\beta_k^\pm + i} \end{pmatrix}. \tag{3.5}$$

Inserting (3.5) into (3.3) and equating coefficients one obtains the following linearly independent solutions:

$Y_k^+(r)$ by choosing $\alpha_k^+ \equiv 0, \beta_k^+ \equiv \rho_k, a_0^+(k) \equiv 1, b_0^+(k) \equiv 0$ and for $i \ge 0$,

$$\text{(3.6a)} \qquad \begin{aligned} a_{i+1}^+(k) &\equiv \left[\sum_{j=0}^{i} g_{i+1-j}(d_1(k)a_j^+(k) + d_2(k)b_j^+(k))\right] \Big/ (i+1), \\ b_{i+1}^+(k) &\equiv \left[\sum_{j=0}^{i} g_{i+1-j}(d_3(k)a_j^+(k) + d_4(k)b_j^+(k))\right] \Big/ (i+1+\rho_k); \end{aligned}$$

if ρ_k is not an integer:

$Y_k^-(r)$ by choosing $\alpha_k^- \equiv -\rho_k$, $\beta_k^- \equiv 0$, $a_0^-(k) \equiv 0$, $b_0^-(k) \equiv 1$ and for $i \ge 0$,

$$\text{(3.6b} \qquad \begin{aligned} a_{i+1}^-(k) &\equiv \left[\sum_{j=0}^{i} g_{i+1-j}(d_1(k)a_j^-(k) + d_2(k)b_j^-(k))\right] \Big/ (i+1-\rho_k), \\ b_{i+1}^-(k) &\equiv \left[\sum_{j=0}^{i} g_{i+1-j}(d_3(k)a_j^-(k) + d_4(k)b_j^-(k))\right] \Big/ (i+1). \end{aligned}$$

Then we obtain two linearly independent solutions $r^{\lambda_k^+}X_k^+(r)$ and $r^{\lambda_k^-}X_k^-(r)$ to (2.17a, b), where $r^{\lambda_k^\pm}X_k^\pm(r) \equiv \mathcal{X}_k(r)Y_k^\pm(r)$. Hence the general solution is given by (3.2a) with $c = 0$ and $X_k^{\pm}(r)$ defined by (3.2e) with $\hat{a}_k^{\pm}(0) + \hat{b}_k^{\pm}(0) \equiv 1$. Furthermore, if $C(k)$ and M are such that $M \geq 1 + 2^m$ and $|a_i^{\pm}(k)|$, $|b_i^{\pm}(k)| \leq C(k)M^i$ for all $i < 2\mathcal{G}\mathcal{D}(|k| + n) + \rho_k$ it follows from (3.6), (3.5), and Lemma 3.1 that $|a_i^{\pm}(k)|$, $|b_i^{\pm}(k)| \leq C(k)M^i$ for all $i \leq 0$.

If ρ_k is an integer, say $l \geq 1$, then the solution (3.6b) is no longer valid. Taking $Y_k^{+}(r)$ as in (3.6a) to be one solution to (3.3), a routine calculation yields that a second solution, linearly independent to $Y_k^{+}(r)$, is $Y_k^{-}(r) + c\ln(r)Y_k^{+}(r)$, where

$$c \equiv \sum_{j=0}^{l-1} g_{l-j}(d_1(k)a_j^{+}(k) + d_2(k)b_j^{+}(k))$$

and $Y_k^{-}(r)$ is given by (3.5) with $\alpha_k^{-} \equiv -\rho_k \equiv -l$, $\beta_k^{-} \equiv 0$, $a_0^{-}(k) \equiv 0$, $b_0^{-}(k) \equiv 1$. For $i = 0 \rightarrow l-2$, $a_{i+1}^{-}(k)$ is defined by (3.6b), $a_{l-1}^{-}(k) \equiv 0$, for $i = 0 \rightarrow l-1$, $b_{i+1}^{-}(k)$ is defined by (3.6b), and finally for $i \geq l$,

(3.7)
$$\begin{aligned} a_{i+1}^{-}(k) &\equiv \left[-ca_{i+1-l}^{+}(k) + \sum_{j=0}^{i} g_{i+1-j}(d_1(k)a_j^{-}(k) + d_2(k)b_j^{-}(k)) \right] \Big/ (i+1-l), \\ b_{i+1}^{-}(k) &\equiv \left[-cb_{i+1-l}^{+}(k) + \sum_{j=0}^{i} g_{i+1-j}(d_3(k)a_j^{-}(k) + d_4(k)b_j^{-}(k)) \right] \Big/ (i+1). \end{aligned}$$

We obtain two linearly independent solutions $r^{\lambda_k^{+}}X_k^{+}(r)$ and $r^{\lambda_k^{-}}[X_k^{-}(r) + cr^{\rho_k} \ln(r)X_k^{+}(r)]$ to (2.17 a, b) with c possibly nonzero, by setting $r^{\lambda_k^{\pm}}X_k^{\pm}(r) \equiv \mathcal{X}_k(r)Y_k^{\pm}(r)$. It follows that $X_k^{\pm}(r)$ are given by (3.2e) with $\hat{a}_k^{\pm}(0) + \hat{b}_k^{\pm}(0) \equiv 1$. Similarly to the above we can choose constants $C(k)$ and $M \geq 1 + 2^m$, dependent on p and n, such that $|a_i^{\pm}(k)|$, $|b_i^{\pm}(k)| \leq C(k)M^i$ for all $i \geq 0$. □

With the help of Lemma 3.1 and the following corollary to Theorem 3.1 we will establish our main regularity result for (2.3) in the next section.

COROLLARY 3.1. *Let $p \in (1,2) \cup (2,\infty)$ and n be a positive integer. Suppose that $\hat{g}(r) \equiv \sum_{i=1}^{\infty} g_i r^i$ near $r = 0$ and that $|g_i| \leq \eta^i$ for all $i \geq 1$, where $\eta \leq 1$. Then for any integer k, except $k = -n$, the general solution $X_k(r)$ of* (2.17 a, b) *near $r = 0$ is of the form* (3.2). *In addition for $k \geq n$ and for $\eta \leq 1/(8\mathcal{D})$ we have that*

(3.8a)
$$|a_i^{+}(k)|, |b_i^{+}(k)| \leq (8\mathcal{D}\eta k)^i/i!, \qquad i = 0 \rightarrow k,$$

(3.8b)
$$|a_i^{+}(k)|, |b_i^{+}(k)| \geq (8\mathcal{D}\eta k)^k/k!, \qquad i \geq k,$$

where $\mathcal{D} > 1$ is the constant, independent of p, n, and k, appearing in (3.4).

Proof. Clearly Theorem 3.1 applies with the given assumptions on $\hat{g}$. Let $M \equiv 8\mathcal{D}\eta k$. Firstly, we note that for any integer $j \in [1,k]$,

(3.9)
$$I_j \equiv \sum_{i=0}^{j-1} \eta^{j-i}M^i/i! \leq M^j/j!.$$

Equation (3.9) holds for $j = 1$ as $\eta \leq 1$. Assume it is true for $1 \leq j \leq l < k$, then $I_{l+1} \equiv \eta(I_l + M^l/l!) \leq 2\eta M^l/l! \leq M^{l+1}/(l+1)!$. Hence, by induction (3.9) holds for $1 \leq j \leq k$.

We now show that the improved bounds on the coefficients, (3.8), hold. (3.8a) holds for $i = 0$. Assume it is true for $0 \leq i \leq l < k$; then from (3.6a), (3.4), and (3.9) we have that

$$|a_{l+1}^{+}(k)|, |b_{l+1}^{+}(k)| \leq 4\mathcal{D}k \left[\sum_{j=0}^{l} \eta^{l+1-j} M^j / j! \right] \Big/ (l+1)$$

$$\leq M \left[I_{l-1} + M^l/1! \right] / \, 2(l+1) \leq M^{l+1}/(l+1)!.$$

Hence, by induction (3.8a) holds. Inequality (3.8b) holds for $i = k$. Assume it is true for $k \leq i \leq l$. Then as $M \leq k$ we have from (3.6a), (3.4), (3.9), and (3.8a) that

$$|a_{l+1}^{+}(k)|, |b_{l+1}^{+}(k)| \leq [8\mathcal{D}\eta k/(l+1)] \left[\sum_{j=0}^{k-1} \eta^{l-j} M^j/j! + M^k/k! \sum_{j=k}^{l} \eta^{l-j} \right] \Big/ 2$$

$$\leq \left[\eta^{l-k} + \sum_{j=0}^{l-k} \eta^j \right] (M^k/k!)/2 \leq (M^k/k!).$$

Hence, by induction (3.8b) holds. □

4. The main regularity result.

THEOREM 4.1. *Let $u \in W_{\text{loc}}^{1,q}(\Omega)$, $1 < q < \infty$, be a local weak solution of* (2.3) *with φ satisfying* (A1)–(A3). *Then*

$$u \in C^{k,\beta}(\Omega) \cap W_{\text{loc}}^{k+2,v}(\Omega), \tag{4.1}$$

where the integer $k \geq 1$ and exponents $\beta \in (0,1]$ and $v \in [1,2]$ are determined uniquely from

$$\begin{aligned} &6(k+\beta) = 7 + 1/(p-1) + [1 + 14/(p-1) + 1/(p-1)^2]^{1/2} \\ &\text{and} \quad 1 \leq v < 2/(2-\beta), \end{aligned} \tag{4.2}$$

where $p \equiv 2 + \Gamma(0)$.

Proof. Let $\hat{g}$ be defined as in (2.17c). From the definition of G in Lemma 2.1, Assumption (A3), and Remark 2.1 we can assume that $\hat{g}(r) \equiv \sum_{i=1}^{\infty} g_i r^i$ near $r = 0$ with $|g_i| \leq \eta^i$ for all $i \geq 1$ with $\eta > 0$ as small as we please. Hence we can apply Corollary 3.1 to the system (2.17), yielding that the general solution has the form (3.2a) about $r = 0$ if $k \neq -n$. We note that in this case one can obtain $X_{-n}(r)$ directly from $X_n(r)$. For all integers k the constant C_k^- in (3.2a) must be set equal to zero since $\lambda_k^- < 0$, $\hat{a}_k^{\pm}(0) + \hat{b}_k^{\pm}(0) \equiv 1$ and we know that $X_k \in C[0,\delta]^2$; see (2.15) and (2.17). For the same reason $C_k^+ \equiv 0$ when $|k| < n$ since $\lambda_k^+ < 0$. In addition, as $H(0) = 0$, $\lambda_n^+ = 0$, $\hat{a}_n^+(0) = 1$, and $\hat{b}_n^+(0) = 0$ it follows that $C_n^+ = 0$. Therefore we have from (3.6) and (2.17) that

$$\begin{aligned} H(r,\theta) &\equiv \sum_{k\geq|n+1|} C_k^+ r^{\lambda_k^+}(\hat{a}_k^+(r)+\hat{b}_k^+(r))e^{i(k-n)\theta} \\ &\equiv \sum_{k=n+1}^{\infty}\Big[\Big(C_k^+e^{ik\theta}+\varepsilon_k^+\overline{C_k^+}e^{-ik\theta}\Big)\hat{a}_k^+(r) \\ &\qquad +(C_k^+e^{ik\theta}+\varepsilon_k^-\overline{C_k^+}e^{-ik\theta})\hat{b}_k^+(r)\Big]r^{\lambda_k^+}e^{-in\theta} \end{aligned} \tag{4.3a}$$

or equivalently,

$$\begin{aligned} H(\xi) \equiv \sum_{k=n+1}^{\infty}&\Big[\Big(C_k^+\xi^k+\varepsilon_k^+\overline{C_k^+}\bar{\xi}^k\Big)\hat{a}_k^+(|\xi|) \\ &+\Big(C_k^+\xi^k+\varepsilon_k^-\overline{C_k^+}\bar{\xi}^k\Big)\hat{b}_k^+(|\xi|)\Big]|\xi|^{\lambda_k^++n-k}\xi^{-n}. \end{aligned} \tag{4.3b}$$

Since $H\in L^2(B)$ we have from (4.3) that for R sufficiently small,

$$\begin{aligned} &\int_0^R r\int_0^{2\pi}|H(r,\theta)|^2 d\theta dr \\ &\equiv 2\pi\sum_{k=n+1}^{\infty}|C_k^+|^2\int_0^R r^{2\lambda_k^++1}[(\hat{a}_k^+(r)+\hat{b}_k^+(r))^2 \\ &+(\varepsilon_k^+\hat{a}_k^+(r)+\varepsilon_k^-\hat{b}_k^+(r))^2]dr<\infty. \end{aligned} \tag{4.4}$$

In the analysis that follows M_i denotes a constant dependent only on $R<1$ and p. We set $\sigma_i\equiv a_i^+(k)+b_i^+(k)$, $t=1/r$ and note from (3.2b) that $\lambda_k^+\leq k(p-1)^{1/2}$. In addition we note that the L^1 norm over an interval of length 2 of a polynomial of degree k with leading coefficient 1 is bounded below by 2^{1-k}; see, for example, p. 80 in [25]. Then for $k\geq n+1$ we have from (3.8) that

$$\begin{aligned} &R\left[\int_0^R r^{2l_k^++1}[\hat{a}_k^+(r)+\hat{b}_k^+(r)]^2dr\right]^{1/2} \\ &\geq\int_0^R r^{\lambda_k^++1}|\hat{a}_k^+(r)+\hat{b}_k^+(r)|dr \\ &\geq\int_0^R r^{\lambda_k^++1}\left[\left|\sum_{i=0}^{k}\sigma_i r^i\right|-2[(8\mathcal{D}\eta k)^k/k!]\sum_{i=k+1}^{\infty}{}' r^i\right]dr \\ &\geq\int_{1/R}^{\infty}t^{-(\lambda_k^++k+3)}\left|\sum_{i=0}^{k}\sigma_i t^{k-i}\right|dt-(M_1\eta k)^k/k! \end{aligned} \tag{4.5}$$

$$\geq M_2^k \int_{1/R}^{(2R+1)/R} \left| \sum_{i=0}^{k} \sigma_i t^{k-i} \right| dt - (M_1 \eta k)^k / k!$$

$$\geq M_3^k - (M_1 \eta k)^k / k! \geq M_3^k - (M_1 \eta e)^k,$$

since $\sigma_1 \equiv 1$ and $(k/e)^k/k! \leq e^{-1}$. Combining (4.4) and (4.5) we have that by choosing η sufficiently small there exists an M, dependent only on R and p, such that

$$|C_k^+| \leq M^k \quad \text{for all } k \geq n+1. \tag{4.6}$$

Let δ be such that δ, $Me\delta^{(p-1)^{1/2}} \in (0, R)$. As $\lambda_k^+ \geq k(p-1)^{1/2} - pn/2$ we have that $\sum_{k=n+1}^{\infty} (Me)^k \delta^{\lambda_k^+} < \infty$. It then follows from (4.3b), (4.6), (3.2d), and (3.8) that for $|\xi| \leq \delta$,

$$\begin{aligned} |H(\xi)| &\leq \sum_{k=n+1}^{\infty} |C_k^+| \left[(1 + |\varepsilon_k^+|)|\hat{a}_k^+(|\xi|)| + (1 + |\varepsilon_k^-|)|\hat{b}_k^+(|\xi|)| \right] |\xi|^{\lambda_k^+} \\ &\leq C(|\xi|/\delta)^{\lambda_{n+1}^+} \sum_{k=n+1}^{\infty} [|\hat{a}_k^+(|\xi|)| + |\hat{b}_k^+(|\xi|)|] M^k \delta^{\lambda_k^+} \\ &\leq C|\xi|^{\lambda_{n+1}^+} \sum_{k=n+1}^{\infty} (Me)^k \delta^{\lambda_k^+} \leq C|\xi|^{\lambda_{n+1}^+}. \end{aligned} \tag{4.7}$$

Therefore the upper bound of (41) in [13] holds. In addition, we see from (4.7) that the series (4.3b) is uniformly and absolutely convergent for $|\xi| \leq \delta$.

As H is a one-to-one mapping near $\xi = 0$ it follows that $|C_{n+1}^+| \neq 0$. Noting from (3.2d) that $|\varepsilon_{n+1}^+| < 1$, it follows from (4.3b), (4.7), and (3.8) that there exist positive constants M_i such that for $|\xi|$ sufficiently small,

$$\begin{aligned} |H(\xi)| &\geq |C_{n+1}^+||\xi|^{\lambda_{n+1}^+} \left[M_1 |\hat{a}_{n+1}^+(|\xi|)| - M_2 |\hat{b}_{n+1}^+(|\xi|)| \right] - M_3 |\xi|^{\lambda_{n+2}^+} \\ &\geq |C_{n+1}^+| M_1 |\xi|^{\lambda_{n+1}^+} \left[1 - M_4|\xi| - M_5 |\xi|^{\lambda_{n+2}^+ - \lambda_{n+1}^+} \right] \geq |C_{n+1}^+| M_6 |\xi|^{\lambda_{n+1}^+}. \end{aligned}$$

Hence, the lower bound of (41) in [13] holds. Similarly, one can prove that the bounds (42) and (43) in [13] hold. The remainder of the proof of (4.1) is then identical to that in [13]; that is, apply Lemmas 1 and 2 and Corollary 2 in [13]. □

Note that the conclusion of Theorem 4.1 is not necessarily true for (1.1) for $\rho \neq 0$ (even if ρ is a constant); see [29].

Example 4.1. For the examples considered in Example 2.1 we have with $q > 1$ that the local weak solutions of (2.3) are such that (4.1) and (4.2) hold with $p \equiv 2 + \Gamma(0)$ given as follows: (i) $\varphi(t) \equiv t^{q-2} \Rightarrow p = q$; (ii) $\varphi(t) \equiv [t(1+t)]^{(q-2)/2} \Rightarrow p = (q+2)/2$; and (iii) $\varphi(t) \equiv [t(1+t^2)]^{(q-2)/3} \Rightarrow p = (q+4)/3$. Therefore we see that the solutions of (2.3) with φ defined by (ii) and (iii) are as regular as those of (2.3) with $\varphi(t) \equiv t^{p-2}$, which is nothing other than the singular part of φ in (ii) and (iii). In other words, it is the singular part of φ that determines the regularity of the local solutions of (2.3).

5. A global regularity result. As mentioned in §1, in order to obtain explicit error bounds for the finite element approximation of (2.3) it is necessary to extend the local regularity results obtained in the last section into global results. Apart from those global $C^{1,\alpha}$ and H^2 results mentioned in the introduction, there appear to be no other global results in the literature. We derive in this section a global regularity result from Theorem 4.1 under some restrictions on the boundary data.

Consider the following mechanics problem. Suppose a plane Ω, which is made of a power law material, is undergoing longitudinal shear (the mode III problem; see [3]). In this case the only nonvanishing component of displacement is the z-component $u(x,y)$, which satisfies (2.3) with $\varphi(t) \equiv t^{p-2}$ and $p \in (1,2)$. Physically one expects that the tangential strain of the plane along $\partial\Omega$ will vary strictly monotonically except at those points where the displacement $u(x,y)$ achieves its extreme values. Motivated by this observation we will suppose that (2.3) is supplemented by the boundary condition

$$u|_{\partial\Omega} = g|_{\partial\Omega}, \tag{5.1}$$

where g has a nonvanishing tangential derivative, g_τ, on $\partial\Omega$ except at those points where g achieves its global extrema on $\partial\Omega$. We have the following result.

THEOREM 5.1. *Let φ satisfy* (A1) *with $C^\infty(0,\infty)$ replaced by $C^{1,1}(0,\infty)$,* (A2) *and* (A3) *with $p \equiv 2 + \Gamma(0) \in (1,2)$. Let $\partial\Omega \in C^{2,1}$ and $g \in C^1(\bar{\Omega})$ be such that $|g_\tau(x,y)| > 0$ if $(x,y) \in \partial\Omega$ and $g(x,y)$ is not a global extremum of g on $\partial\Omega$. Let $u \in W^{1,q}(\Omega)$ be the unique solution of* (2.3) *with the boundary condition* (5.1). *Then*

$$\text{(i)} \qquad g \in C^{2,\gamma}(\bar{\Omega}), \qquad \gamma > 0, \Rightarrow u \in C^{2,\mu}(\bar{\Omega}),$$

$$\text{(ii)} \qquad g \in C^{2,\gamma}(\bar{\Omega}) \cap W^{3,r}(\Omega), \qquad \gamma > 0 \text{ and } r \in (1,p], \Rightarrow u \in C^{2,\mu}(\bar{\Omega}) \cap W^{3,r}(\Omega),$$

where $\mu = \gamma$ for $p \in (1,1.3), \mu = \min(\gamma,\beta)$ for $p \in [1.3,2)$, and

$$\beta = \{-5 + 1/(p-1) + [1 + 14/(p-1) + 1/(p-1)^2]^{1/2}\}/6. \tag{5.2}$$

Proof. From [19] we have that $g \in C^{1,\alpha}(\bar{\Omega}) \Rightarrow u \in C^{1,\alpha}(\bar{\Omega})$ for some $\alpha > 0$, so that $|\nabla u|$ is continuous on $\bar{\Omega}$. We first show that $|\nabla u|_{\partial\Omega}| \geq \delta > 0$. If $(x,y) \in \partial\Omega$ and $g_\tau(x,y) \neq 0$, then $|\nabla u(x,y)| > 0$ as $(u-g)|_{\partial\Omega} = 0$. Let $g_\tau(x_0,y_0) = 0$ and (without loss of generality) let $g(x,y) \geq g(x_0,y_0)$ for all $(x,y) \in \partial\Omega$. Let $u_0 \equiv g(x_0,y_0)$. Then $u - u_0$ is such that $\nabla.(\varphi(|\nabla(u-u_0)|)\nabla(u-u_0)) \equiv \nabla.(\varphi(|\nabla u|)\nabla u) = 0$ in Ω and $(u-u_0)_{\partial\Omega} \geq 0$ because u_0 is a constant. If u is constant, then the desired regularity result holds. Therefore we can assume that u is not constant. From the strong maximum principle in §3 of [28], it follows that $u > u_0$ in Ω. (This result also holds from the proof of the strong comparison principle in [22].) From the Hopf maximum principle, see [28] and Lemma 3.4 in [12], we have that $|\partial u/\partial n(x_0,y_0)| > 0$, where $\underline{n}$ is the outward unit normal to $\partial\Omega$. Consequently, $|\nabla u|_{\partial\Omega}| > 0$ and from the continuity of $|\nabla u|$, there exists a domain $S \subset\subset \Omega$ such that $\partial S \in C^{2,1}$ and $|\nabla u|_{\bar{\Omega}\setminus S}| \geq \delta > 0$.

Let $S \subset\subset S_1 \subset\subset \Omega$. Then from Theorem 4.1 we have that

$$p \in (1, 1.3) \Rightarrow k \geq 3 \Rightarrow u \in C^3(\overline{S_1}), \tag{5.3a}$$

$$p \in [1.3, 2) \Rightarrow k = 2 \Rightarrow u \in C^{2,\beta}(\overline{S_1}) \cap W^{3,p}(S_1) \quad \text{with } \beta\text{given by (5.2).} \tag{5.3b}$$

Rewriting (2.3), we have for $Q \equiv \Omega \backslash S$ and $g \in C^{2,\gamma}(\bar{\Omega})$ that

$$\Delta u + \Gamma(|\nabla u|) \left[(u_x)^2 u_{xx} + 2u_x u_y u_{xy} + (u_y)^2 u_{yy}\right] / |\nabla u|^2 = 0 \quad \text{in } Q,$$

$$u|_{\partial Q} \in C^{2,\mu}(\partial Q).$$

As $\Gamma(t) \equiv [(\varphi(t)t)'/\phi(t)] - 1 \geq q - 2$, it follows from our assumptions that $\Gamma \in C^{0,1}(0, \infty)$ and hence the coefficients $\Gamma(|\nabla u|)(u_x)^2/|\nabla u|^2$, $\Gamma(|\nabla u|)(u_x u_y)/|\nabla u|^2$, and $\Gamma(|\nabla u|)(u_y)^2/|\nabla u|^2 \in C^{0,\alpha}(\bar{Q})$. It then follows from Theorem 9.19 in [12], for example, that $u \in C^{2,\sigma}(\bar{Q})$, $\sigma = \min(\alpha, \mu)$. This yields that the above coefficients then belong to $C^{0,1}(\bar{Q})$ and hence $u \in C^{2,\mu}(\bar{Q})$. Combining this with (5.3) yields that $u \in C^{2,\mu}(\bar{\Omega})$. If $g \in C^{2,\alpha}(\bar{\Omega}) \cap W^{3,r}(\Omega)$, $r \in (1, p]$ it follows in a similar manner that $u \in C^{2,\alpha}(\bar{\Omega}) \cap W^{3,r}(\Omega)$. □

It is of great interest to know if the conclusion of Theorem 5.1 is true or not for more general bounary conditions. Of course, it would be even better if Theorem 4.1 could be extended to a global result. Nevertheless, the above result shows that the global regularity on u required to guarantee the optimal rate of convergence in $W^{1,p}(\Omega)$ of the continuous piecewise linear finite element approximation of (2.3) does hold for suitable data. Another unsolved problem is extending the sharp local regularity results in [13] to higher dimensions.

REFERENCES

[1] G. Aronsson, *Representation of a p-harmonic function near a critical point in the plane*, Manuscripta Math., 66 (1989), pp. 73–95.

[2] G. Aronsson and P. Lindqvist, *On p-harmonic functions in the plane and their stream functions*, J. Differential Equations, 74 (1988), pp. 157–178.

[3] C. Atkinson and C. R. Champion, *Some boundary value problems for the equation* $\nabla .(|\nabla\phi|^N \nabla\phi) = 0$, Quart. J. Mech. Appl. Math., 37 (1984), pp. 401–419.

[4] C. Atkninson and C. W. Jones, *Similarity solutions in some non-linear diffusion problems and in boundary-layer flow of a pseudo plastic fluid*, Quart. J. Mech. Appl. Math., 27 (1974), pp. 193–211.

[5] J. Baranger and K. Najib, *Numerical analysis for quasi-Newtonian flow obeying the power or the Carreau law*, Numer. Math., 58 (1990), pp. 35–49.

[6] J. W. Barrett and W. B. Liu, *Finite element approximation of the p-Laplacian*, Math. Comp., to appear.

[7] B. Bojarski and T. Iwaniec, *p-harmonic equation and quasi regular mappings*, in Partial Differential Equations, Banach Center Publications 19, PWN-Polish Scientific Publishers, Warsaw, 1987, pp. 25–38.

[8] S. S. Chow, *Finite element error estimates for non-linear elliptic equations of monotone type*, Numer. Math., 54 (1989), pp. 373–393.

[9] E. DiBenedetto, *$C^{1+\alpha}$ local regularity of weak solutions of degenerate elliptic equations*, Nonlinear Anal., 7 (1983), pp. 827–850.

[10] E. DiBenedetto and A. Friedman, *Regularity of solutions of nonlinear degenerate parabolic systems*, J. Reine Angew. Math., 349 (1984), pp. 83–128.

[11] M. Giaquinta, *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Systems*, Princeton Univ. Press, Princeton, NJ, 1983.

[12] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Second ed., Springer-Verlag, Berlin, 1984.

[13] T. Iwaniec and J. J. Manfredi, *Regularity of p-harmonic functions on the plane*, Rev. Math. Iberoamericana, 5 (1989), pp. 1–19.

[14] A. Kufner, O. John, and S. Fučik, *Function Spaces*, Nordhoff, Leyden, 1977.

[15] O. A. Ladyzhenskaya and N. N. Ural'tseva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.

[16] O. Lehto and K. Virtanen, *Quasiconformal Mappings in the Plane*, Springer-Verlag, Berlin, 1973.

[17] J. L. Lewis, *Smoothness of certain degenerate elliptic equations*, Proc. Amer. Math. Soc., 80 (1980), pp. 259–265.

[18] ———, *Regularity of the derivatives of solutions to certain elliptic equations*, Indiana Univ. Math. J., 32 (1983), pp. 849–858.

[19] G. M. Lieberman, *Boundary regularity for solutions of degenerate elliptic equations*, Nonlinear Anal., 12 (1988), pp. 1203–1219.

[20] W. B. Liu and J. W. Barrett, *A remark on the regularity of the solutions of the p-Laplacian and its applications to their finite element approximation*, J. Math. Anal. Appl., to appear.

[21] ———, *A further remark on the regularity of the solutions of the p-Laplacian and its applications to their finite element approximation*, Nonlinear Anal., to appear.

[22] J. J. Manfredi, *p-harmonic functions in the plane*, Proc. Amer. Math. Soc., 103 (1988), pp. 473–479.

[23] C. B. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, Berlin, 1966.

[24] J. R. Philip, *N-diffusion*, Austral. J. Phys., 14 (1961), pp. 1–13.

[25] T. J. Rivlin, *The Chebyshev Polynomials*, John Wiley, New York, 1974.

[26] F. De Thelin, *Local regularity properties for the solutions of a nonlinear partial differential equation*, Nonlinear Anal., 6 (1982), pp. 839–844.

[27] P. Tolksdorf, *On quasilinear boundary value problems in domains with corners*, Nonlinear Anal., 5(1981), pp. 721–735.

[28] ———, *On the Dirichlet problem for quasilinear equations in domains with conical boundary points*, Comm. Differential Eqns., 8 (1983), pp. 773–817.

[29] ———, *Regularity for a more general class of quasilinear elliptic equations*, J. Differential Eqns., 51 (1984), pp. 126–150.

[30] K. Uhlenbeck, *Regularity for a class of nonlinear elliptic systems*, Acta. Math., 138 (1977), pp. 219–240.

[31] N. N. Ural'tseva, *Degenerate quasilinear elliptic systems*, Zap. Nauchno. Sem. Leningrad. Otdel. Mat. Inst. Steklov., 7 (1968), pp. 184–222. (In Russian.)

SIAM J. MATH. ANAL.
Vol. 24, No. 6, pp. 1537–1556, November 1993

007

GLOBAL OPTIMALITY CRITERION AND A DUALITY WITH A ZERO GAP IN NONCONVEX OPTIMIZATION[*]

PHAN THIEN THACH[†]

Abstract. This paper presents a relation between geometrical criteria for optimality and a duality in two nonconvex problems: a quasi-convex maximization over a convex set, and a general minimization over the complement of a convex set. A connection between the duality by Toland, the global optimality criterion by Hiriart-Urruty, and the author's result is also given. Several applications are presented.

Key words. geometrical criteria, nonconvex duality

AMS subject classifications. 49N15, 49J52, 90C26

1. Introduction.

Duality theories have provided important relationships in both analytical and algorithmic studies on optimization problems. Primal-dual pairs allow us to construct primal-dual algorithms whose efficiency and nice economic interpretation have attracted many researchers. There are several approaches used to obtain dual problems of a given mathematical program. For example, in convex duality theory we can find dual problems by either the Lagrangian functional approach or the perturbation functional approach. These approaches have a common starting point, the Kuhn–Tucker condition. A Kuhn–Tucker vector provides optimal Lagrange multipliers in the first approach and it is a minimizer of the conjugate of the perturbation functional in the second approach.

In convex programs local optimality criteria are also global. In nonconvex programs the local information such as subdifferentials and tangent cones are not enough to provide a global optimality criterion. However, by using the ε-subdifferentials Hiriart-Urruty has provided a global optimality criterion for a problem of minimizing the difference of two convex functions [5]. Since the minimization of the difference of two convex functions can be converted into a convex maximization over a convex set or a convex minimization over the complement of a convex set, a natural question is how to extend the optimality criterion to more general problems. In this paper we present geometrical criteria for global optimality in two problems: a quasi-convex maximization over a closed set and a general minimization over the complement of an open convex set. For constructing optimality criteria the separation theorem (or its variant) is a basic instrument. We say that criterion A is reducible to criterion B if we can obtain criterion B from criterion A by a transformation where the separation theorem is not used. Criterion A is equivalent to criterion B if A is reducible to B and B is reducible to A, i.e., in the proof for the equivalence we do not use the separation theorem. With this definition our geometrical criterion will be equivalent to the criterion in [5] if we consider a problem of minimizing the difference of two convex functions as a particular case of reverse convex constraint problems.

We also present a connection between the optimality criteria and a nonconvex duality. These criteria involve the polars of level sets of the objective functions, and there is a polarity correspondence between the level sets of the objective functions in

[*] Received by the editors October 10, 1991; accepted for publication (in revised form) March 8, 1993. The first manuscript of this paper was prepared when the author was visiting the Department of Mathematics, University of Trier as a fellow of the Alexander von Humboldt Foundation.

[†] Present address: Institute of Human and Social Sciences, Tokyo Institute of Technology. The author is on leave from the Hanoi Institute of Mathematics, Hanoi, Vietnam.

a primal-dual pair. By our approach we can obtain dual problems given in Toland [25], and Hiriart-Urruty [4].

The paper consists of six sections. In §2 we present geometrical criteria for global optimality in general nonconvex problems. In §3 we show a polarity correspondence between the level sets of a function and level sets of its quasi conjugate. In §4 we present a duality scheme. Several applications are given in §5. Finally we draw some conclusions in §6.

2. Geometrical criteria for global optimality in nonconvex problems.

Let X be a closed convex set in R^n. It is well known that a vector x does not belong to X if and only if there is a hyperplane strictly separating x and X. The direction "if" is straightforward, and the direction "only if" is a separation theorem. Now let X_1, X_2 be closed convex sets containing 0. Denote by X_1^0, X_2^0 the polars of X_1, X_2, respectively:

$$X_1^0 = \{v : \langle v, x\rangle \leq 1 \;\; \forall x \in X_1\}, \qquad X_2^0 = \{v : \langle v, x\rangle \leq 1 \;\; \forall x \in X_2\}.$$

From the definition of the polar sets it is straightforward that

$$X_1 \subseteq X_2 \Rightarrow X_2^0 \subseteq X_1^0.$$

Since X_1, X_2 are closed convex and contain zero, by the separation theorem we can prove that

$$(X_1^0)^0 = X_1, \qquad (X_2^0)^0 = X_2.$$

Therefore, we have the inverse direction:

$$X_2^0 \subseteq X_1^0 \Rightarrow X_1 \subseteq X_2.$$

Thus in the equivalent relation

$$X_1 \subseteq X_2 \Leftrightarrow X_2^0 \subseteq X_1^0, \tag{1}$$

the separation theorem plays a crucial role in the direction "$\Leftarrow$". Similarly by the strict separation we can prove (1) if X_1, X_2 are open convex sets containing 0. If X_1 and X_2 are neither closed nor open, then generally speaking we do not have (1) (more precisely, the direction "$\Leftarrow$" in (1)). This is quite reasonable, because we will lose the information of irregular points on the boundary of a convex set if we pass to its polar.

In this section we use relation (1) to interpret, in the dual space, the optimality for a solution of a primal problem. Throughout this paper the terms "normal cone," "polar set," "conjugate function," and others have the meanings as in Rockafellar [14].

We start from a very simple convex program

$$\max\{\langle c, x\rangle : x \in D\}, \tag{2}$$

where c is a nonzero vector in R^n, and D is a closed convex set in R^n such that $0 \in \text{int}D$. It is well known that a vector $x^* \in D$ is optimal if and only if c belongs to the normal cone of D at x^*:

$$c \in N(x^*, D). \tag{3}$$

This is equivalent to $(1/\langle c, x^*\rangle)c \in D^0$. Since $0 \in D^0$, this is equivalent to

$$\begin{aligned}&\left\{\frac{\lambda}{\langle c, x^*\rangle}c : 0 \leq \lambda \leq 1\right\} \subseteq D^0\\ \Leftrightarrow &\left\{\frac{\lambda}{\langle c, x^*\rangle}c : 0 \leq \lambda \leq 1\right\}^0 \supseteq D^{00}\\ \Leftrightarrow &\, D \subseteq \{x : \langle c, x\rangle \leq \langle c, x^*\rangle\}.\end{aligned}$$

Thus, criterion (3) can be interpreted as the use of the relation (1) to problem (2).

We consider now more general problems. The first type of nonconvex problems is

$$(4) \qquad \sup\{f(x) : x \in D\},$$

where f is a quasi-convex function, D is a closed subset such that $0 \in \mathrm{conv}D$ (conv$D$ stands as the convex hull of D). Since f is quasi-convex, the supremum value of f over D is equal to the supremum of f over conv$D$. However, in (4) we do not replace D by conv$D$, because in order to find an optimal solution (sometimes called briefly "solution") we do not need to handle the convex hull of D (see Example 4.1 in §4). By setting $\overline{f}(x) = \max\{f(x), f(0)\}$ we see that a maximizer of $\overline{f}$ on D is also a maximizer of f on D. Therefore, we can assume that

$$f(0) = \min\{f(x) : x \in R^n\}.$$

Using relation (1) we have the following geometrical criterion for the optimality in problem (4).

THEOREM 2.1. *Let $z \in D$. The condition*

$$(5) \qquad \{x : f(x) \leq f(z)\}^0 \subseteq D^0$$

is necessary for the optimality of z. If

$$(6) \qquad \sup\{f(x) : x \in D\} = \sup\{f(x) : x \in \mathrm{int}(\mathrm{conv}D)\},$$

then (5) *is sufficient for the optimality of z.*

Proof. If $z \in D$ is optimal, then $D \subseteq \{x : f(x) \leq f(z)\}$. By definition of the polar one has (5). Now suppose that (6) is satisfied and $z \in D$ is not optimal. Then there is vector x^* such that

$$x^* \in D \setminus \mathrm{cl}\{x : f(x) \leq f(z)\},$$

where cl stands as the closure. Since $0 \in \{x : f(x) \leq f(z)\}$, this implies that there is vector $y \in R^n$ such that

$$\langle y, x^*\rangle > 1 \geq \langle y, x\rangle \quad \forall x : f(x) \leq f(z).$$

The second inequality implies $y \in \{x : f(x) \leq f(z)\}^0$. However, $y \notin D^0$ because

$$\sup\{\langle y, x\rangle : x \in D\} \geq \langle y, x^*\rangle > 1.$$

Thus (5) is not satisfied. □

Criterion (5) is rather straightforward, but it can be used later to obtain a nonconvex duality. If f is lower semicontinuous (lsc) and $\mathrm{int}(\mathrm{conv}D) \neq \emptyset$, then we can check that condition (6) is satisfied. In general cases relation (5) is not sufficient for the optimality because optimal solutions could be irregular points on the boundary of a level set of the objective function. We shall say that problem (4) is *regular* if condition (6) is satisfied.

The second type of nonconvex problems under our consideration is

$$\inf\{f(x) : x \notin \operatorname{int} D\}, \tag{7}$$

where f is an arbitrary function and D is a closed convex set containing 0 in its interior.

THEOREM 2.2. *Let $z \notin \operatorname{int} D$. The following condition*

$$D^0 \subseteq \{x : f(x) < f(z)\}^0 \tag{8}$$

is necessary for the optimality of z. If

$$\inf\{f(x) : x \notin \operatorname{int} D\} = \inf\{f(x) : x \notin D\}, \tag{9}$$

then condition (8) *is sufficient for the optimality of z.*

Proof. Suppose that z is optimal to problem (7). Condition (8) follows immediately from $\{x : f(x) < f(z)\} \subseteq \operatorname{int} D \subseteq D$. Suppose now that $z \notin \operatorname{int} D$ is not optimal and condition (9) is satisfied. Then there is $u \notin D$ such that $f(z) > f(u)$. Since $u \notin D$ and D is closed and contains 0, there is a vector $v \in D^0$ such that $\langle u, v\rangle > 1$. Since $f(u) < f(z)$, this implies $v \notin \{x : f(x) < f(z)\}^0$. Therefore, $D^0 \not\subseteq \{x : f(x) < f(z)\}^0$. □

We shall say that problem (7) is regular if condition (9) is satisfied. It is easy to see that if f is upper semicontinuous (usc), then problem (7) is regular.

An important particular case of problem (7) is

$$\inf_{x \in R^n} \{h_0(x) - g_0(x)\}, \tag{10}$$

where $h_0(\cdot)$ is an extended real valued function and g_0 is finite convex function. Using the ε-subdifferentials Hiriart-Urruty [5] proved the following criterion for the optimality of a vector z:

$$\partial_\varepsilon g_0(z) \subseteq \partial_\varepsilon h_0(z) \quad \forall \varepsilon > 0. \tag{11}$$

We know that by an additional variable we can convert problem (10) into a minimization over the complement of a convex set. Therefore, it is natural to study the connection between criteria (8) and (11). In the sequel we prove that these criteria are equivalent for the same problem.

Set

$$h_1(x) = h_0(z + x) - h_0(z), \qquad g_1(x) = g_0(z + x) - g_0(z).$$

Problem (10) is converted into

$$\inf\{f(x, t) : (x, t) \notin \operatorname{int} D\}, \tag{12}$$

where

$$f(x, t) = h_1(x) - t, \qquad D = \{(x, t) : g_1(x) - t \leq 0\}.$$

Problem (12) is regular because

$$\inf\{h_1(x) - t : g_1(x) - t \geq 0\} = \inf\{h_1(x) - t : g_1(x) - t > 0\}.$$

Vector z is optimal to problem (10) if and only if vector $(0, g_1(0)) \in R^n \times R$ is optimal to problem (12). Since $f(0, g_1(0)) = 0$, criterion (8) at $(0, g_1(0))$ is

$$\{(x,t) : g_1(x) \leq t\}^0 \subseteq \{(x,t) : h_1(x) < t\}^0. \tag{13}$$

PROPOSITION 2.3. *Let h be an arbitrary function defined on R^n. For every $\alpha \in R$, one has*

$$\begin{aligned}&\{(x,t) : h(x) - t - \alpha < 0\}^0\\ &\quad = \{(x,t) : h(x) - t - \alpha \leq 0\}^0\\ &\quad = \{(y,\lambda) : \lambda < 0, h^*(-y/\lambda) + \alpha \leq -1/\lambda\} \cup \{(y,0) : y \in (\operatorname{dom} h)^0\}.\end{aligned}$$

Proof. By definition,

$$\begin{aligned}&(y,\lambda) \in \{(x,t) : h(x) - t - \alpha < 0\}^0\\ &\quad \Leftrightarrow \sup\{\langle y, x\rangle + \lambda.t : h(x) - t - \alpha < 0\} \leq 1\\ &\quad \Leftrightarrow s(y,\lambda) := \sup\{\langle y, x\rangle + \lambda.t : h(x) - t - \alpha \leq 0\} \leq 1\\ &\quad \Leftrightarrow (y,\lambda) \in \{(x,t) : h(x) - t - \alpha \leq 0\}^0.\end{aligned}$$

If $\lambda > 0$ then $s(y,\lambda) = +\infty$. If $\lambda = 0$, $y \notin (\operatorname{dom} h)^0$, then $s(y,\lambda) > 1$. If $\lambda = 0$, $y \in (\operatorname{dom} h)^0$, then $s(y,\lambda) \leq 1$. If $\lambda < 0$, then

$$\begin{aligned}s(y,\lambda) &= \sup_{x \in R^n}\{\langle y, z\rangle + \lambda.(h(x) - \alpha)\} \leq 1\\ &\Leftrightarrow \sup_{x \in R^n}\{\langle -y/\lambda, x\rangle - h(x) + \alpha\} \leq -1/\lambda\\ &\Leftrightarrow h^*(-y/\lambda) + \alpha \leq -1/\lambda.\end{aligned}$$

This completes the proof. □

From Proposition 2.3 one has

$$\{(x,t) : g_1(x) \leq t\}^0 = \{(y,\lambda) : \lambda < 0, g_1^*(-y/\lambda) \leq -1/\lambda\} \cup \{0\} \tag{14}$$

$$\begin{aligned}\{(x,t) : h_1(x) < t\}^0 = \{(y,\lambda) : \lambda < 0, h_1^*(-y/\lambda) \leq -1/\lambda\}\\ \cup\{(y,0) : y \in (\operatorname{dom} h)^0\}.\end{aligned} \tag{15}$$

From (14)–(15) it follows that criterion (13) is equivalent to

$$\begin{aligned}&\{y : g_1^*(0) \leq \varepsilon\} \subseteq \{y : h_1^*(0) \leq \varepsilon\} \quad \forall\varepsilon > 0\\ &\quad \Leftrightarrow \partial_\varepsilon g_1^*(0) \subseteq \partial_\varepsilon h_1^*(0) \quad \forall\varepsilon > 0\\ &\quad \Leftrightarrow \partial_\varepsilon g_0(z) \subseteq \partial_\varepsilon g_1(z) \quad \forall\varepsilon > 0.\end{aligned}$$

Thus, we arrive at criterion (11). Note that in the above arguments we do not use the separation theorem.

3. Quasi-conjugate functionals. We know that any function can be defined via its level sets. Let f be a function. If we denote

$$L(f, \leq, \alpha) = \{x : f(x) \leq \alpha\},$$
$$L(f, <, \alpha) = \{x : f(x) < \alpha\},$$

then f can be defined as follows:

$$\begin{aligned} f(x) &= \inf\{\alpha : x \in L(f, \leq, \alpha)\} \\ &= \sup\{\alpha : x \notin L(f, <, \alpha)\}. \end{aligned}$$

By condition (5) we see that $z \in D$ is optimal to problem (4) if

$$\{L(f, \leq, \alpha)\}^0 \subseteq D^0, \tag{16}$$

where $\alpha = f(z)$. Since $\{L(f, \leq, \alpha)\}^0$ and D^0 are closed, convex sets, if they have nonempty interiors, then (16) is equivalent to

$$\operatorname{int}\{L(f, \leq, \alpha)\}^0 \subseteq \operatorname{int} D^0. \tag{17}$$

If $\operatorname{int}\{L(f, \leq, \alpha)\}^0$ is a level set $L(g, <, \beta)$ of a function g, then (17) implies that

$$g(v) \geq \beta \quad \forall v \notin \operatorname{int} D^0. \tag{18}$$

Therefore, we see intuitively that the problem of finding the smallest level set $L(f, \leq, \alpha)$ containing D (in the primal space) can be interpreted as the problem of finding the biggest level set $L(g, <, \beta)$ contained in $\operatorname{int} D^0$ (in the dual space). In this section we prove that the relationship between f and g is nothing but the quasi conjugation presented in [19].

Now for a given function f we set

$$L_\alpha := (L(f, <, -\alpha))^0, \qquad U_\alpha := \operatorname{int}(L(f, \leq, -\alpha))^0.$$

L_α is closed, convex and U_α is open, convex. If $\alpha \leq \beta$, then $-\alpha \geq -\beta$; hence $L(f, \leq, -\beta) \subseteq L(f, \leq, -\alpha)$ and $L(f, <, -\beta) \subseteq L(f, <, -\alpha)$. This implies

$$L_\alpha \subseteq L_\beta, \qquad U_\alpha \subseteq U_\beta \quad (\alpha \leq \beta). \tag{19}$$

Define f^L and f^U as follows:

$$f^L(y) = \inf\{\alpha : y \in L_\alpha\},$$
$$f^U(y) = \sup\{\alpha : y \notin U_\alpha\}.$$

It is easy to check that

$$\{y : f^L(y) \leq \alpha\} = L_\alpha;$$
$$\{y : f^U(y) < \alpha\} = U_\alpha.$$

Since L_α is closed, convex and U_α is open, convex, function f^L is lsc, quasi-convex and function f^U is usc, quasi-convex. In the sequel, we give an analytical formula for f^L and f^U.

Denote by G^L the set of lsc functions $f : R^n \to R \cup \{\pm\infty\}$ such that $L(f, \leq, \alpha)$ is either a compact set or the whole space R^n, by Q^L the set of functions $f \in G^L$ such that $f(0) = \inf\{f(x) : x \in R^n\}$. The property "$L(f, \leq, \alpha)$ is either a bounded set or the whole space R^n" is equivalent to

$$f(x) \to \sup\{f(x) : x \in R^n\} \quad \text{as} \quad \| x \| \to \infty.$$

Denote by Q^U the set of usc functions $f : R^n \to R \cup \{\pm\infty\}$ such that $f(0) = \inf\{f(x) : x \in R^n\}$.

We call the quasi conjugate of a function $f : R^n \to R \cup \{\pm\infty\}$ the function $f^H : R^n \to R \cup \{\pm\infty\}$ defined as follows:

$$f^H(y) = \begin{cases} -\inf\{f(x) : \langle y, x\rangle \geq 1\} & \text{if } y \neq 0 \\ -\sup\{f(x) : x \in R^n\} & \text{if } y = 0. \end{cases}$$

By definition, it is easy to see that f^H is quasi-convex and

$$f^H(0) = \inf\{f^H(y) : y \in R^n\}.$$

THEOREM 3.1. (i) *If $f \in Q^U$ then $f^L \in Q^L$ and $f^L(y) = f^H(y)$ for all $y \neq 0$;*
(ii) *If $f \in G^L$ then $f^U \in Q^U$ and $f^U(y) = f^H(y)$ for all y.*

Proof. (i) Let $f \in Q^U$. For α such that $L(f, <, \alpha)$ is nonempty one has

$$0 \in L(f, <, \alpha) = \text{int}(L(f, <, \alpha))$$

because $f(0) = \inf\{f(x) : x \in R^n\}$. This implies that $L_{-\alpha} = (L(f, <, \alpha))^0$ is compact. For α such that $L(f, <, \alpha)$ is empty, $L_{-\alpha} = \{\emptyset\}^0 = R^n$. Therefore, $f^L \in G^L$. Since $0 \in (L(f, <, \alpha))^0 = L_{-\alpha}$ (for all α), one has

$$f^L(0) = \inf\{\alpha : 0 \in L_{-\alpha}\} = -\infty \leq f^L(y) \quad \forall y.$$

So $f^L \in Q^L$. In order to prove $f^L(z) = f^H(z)$ for all $z \neq 0$ we shall prove that

$$\{y : f^L(y) \leq f^H(z)\} = \{y : f^H(y) \leq f^H(z)\} \tag{20}$$

for all $z \neq 0$. Set $\alpha = f^H(z)$. One has

$$\begin{aligned} f^L(y) \leq \alpha \Leftrightarrow y \in L_\alpha \; &\Leftrightarrow \; y \in (L(f, <, -\alpha))^0 \\ &\Leftrightarrow \sup\{\langle y, x\rangle : f(x) < -\alpha\} \leq 1 \end{aligned} \tag{21}$$

and

$$\begin{aligned} f^H(y) \leq \alpha &\Leftrightarrow -\inf\{f(x) : \langle y, x\rangle \geq 1\} \leq \alpha \\ &\Leftrightarrow \inf\{f(x) : \langle y, x\rangle \geq 1\} \geq -\alpha. \end{aligned} \tag{22}$$

We prove first (21)⇒(22). If (22) is not satisfied then there is vector u such that $\langle y, u\rangle \geq 1$, $f(u) < -\alpha$. Since $\{x : f(x) < -\alpha\}$ is open, one has

$$\sup\{\langle y, x\rangle : f(x) < -\alpha\} > \langle y, u\rangle \geq 1.$$

This conflicts with (21). We prove now (22)⇒(21). If (21) is not satisfied, then there is vector u such that $f(u) < -\alpha$ and $\langle y, u\rangle > 1$. This implies that

$$\inf\{f(x) : \langle y, x\rangle \geq 1\} \leq f(u) < -\alpha.$$

We arrive at the contradiction with (22). Thus, (21) is equivalent to (22) and hence we obtain (20). We would mention that from (22) we obtain (21) without the assumption $f \in Q^U$, and this statement can be regarded as a particular case of more general results in Crouzeix [2].

(ii) Let $f \in G^L$. For every α such that $L(f, \leq, \alpha)$ is compact one has

$$\{y : f^U(y) < -\alpha\} = U_{-\alpha} = \operatorname{int}(L(f, \leq, \alpha))^0 \ni 0. \tag{23}$$

For α such that $L(f, \leq, \alpha) \equiv R^n$ one has

$$\{y : f^U(y) < -\alpha\} = U_{-\alpha} = \operatorname{int}(L(f, \leq, \alpha))^0 = \emptyset. \tag{24}$$

Therefore, f^U is usc and $f^U(0) \leq f^U(y)$ for all $y \in R^n$, i.e., $f^U \in Q^U$. From (23) and (24) it follows that

$$\begin{aligned} f^U(0) &= \sup\{\alpha : 0 \notin U_\alpha\} \\ &= \sup\{\alpha : L(f, \leq, -\alpha) \equiv R^n\} \\ &= -\inf\{\alpha : L(f, \leq, \alpha) \equiv R^n\} \\ &= -\sup\{f(x) : x \in R^n\} \\ &= f^H(0). \end{aligned}$$

In order to prove $f^U(z) = f^H(z)$ for all $z \neq 0$ we shall prove that

$$\{y : f^U(y) < f^H(z)\} = \{y : f^H(y) < f^H(z)\} \tag{25}$$

for all $z \neq 0$. Set $\alpha := f^H(z)$. One has

$$\begin{aligned} f^U(y) < \alpha &\Leftrightarrow y \in U_\alpha \\ &\Leftrightarrow y \in \operatorname{int}(L(f, \leq, -\alpha))^0. \end{aligned} \tag{26}$$

This implies that $L(f, \leq, -\alpha)$ is not the hull space. Since $f \in G^L$, it follows that $L(f, \leq, -\alpha)$ is compact. Therefore, (26) is equivalent to

$$\max\{\langle y, x\rangle : f(x) \leq -\alpha\} < 1. \tag{27}$$

On the other hand,

$$\begin{aligned} f^H(y) &< \alpha \\ &\Leftrightarrow -\inf\{f(x) : \langle y, x\rangle \geq 1\} < \alpha \\ &\Leftrightarrow \inf\{f(x) : \langle y, x\rangle \geq 1\} > -\alpha. \end{aligned} \tag{28}$$

We shall see the equivalence between (27) and (28). First we prove (27)⇒(28). Suppose that (28) is not satisfied, i.e.,

$$\inf\{f(x) : \langle y, x\rangle \geq 1\} \leq -\alpha. \tag{29}$$

Since $f \in G^L$, there is vector u such that $\langle u, x\rangle \geq 1$ and [19]

$$f(u) = \inf\{f(x) : \langle y, x\rangle \geq 1\}.$$

From (29) it follows that $f(u) \leq -\alpha$. Therefore,

$$\sup\{\langle y, x\rangle : f(x) \leq -\alpha\} \geq \langle u, x\rangle \geq 1.$$

This conflicts with (27). Now we prove (28)⇒(27). Suppose that (27) is not satisfied, i.e.,

$$\sup\{\langle y,x\rangle : f(x) \le -\alpha\} \ge 1. \tag{30}$$

Since $L(f,\le,-\alpha)$ is compact, from (30) it follows that there is vector u such that $\langle y,u\rangle \ge 1$ and $f(u) \le -\alpha$. Then,

$$\inf\{f(x) : \langle y,x\rangle \ge 1\} \le f(u) \le -\alpha.$$

This conflicts with (28). Thus, (27)⇔(28) and hence we obtain (25). □

The biquasi-conjugate function $(f^H)^H$ will be denoted by f^{HH}. In [19] we prove that if $f \in Q^U$, then $f^{HH} \in Q^U$ and

$$\{x : f^{HH}(x) < \alpha\} = \operatorname{conv}\{x : f(x) < \alpha\} \quad (\forall\alpha),$$

and that if $f \in Q^L$, then $f^{HH} \in Q^L$ and

$$\{x : f^{HH}(x) \le \alpha\} = \operatorname{conv}\{x : f(x) \le \alpha\} \quad (\forall\alpha).$$

The later relation can be generalized in the following theorem.

THEOREM 3.2. *If $f \in G^L$, then $f^{HH} \in Q^L$ and*

$$\{x : f^{HH}(x) \le \alpha\} = \operatorname{conv}(\{x : f(x) \le \alpha\} \cup \{0\})$$

for all $\alpha > \inf\{f(x) : x \in R^n \setminus \{0\}\}$.

Proof. Let $f \in G^L$. Then, by Theorem 3.1 (ii), $f^H = f^U$ and $f^H \in Q^U$. By Theorem 3.1 (i), $f^{HH}(x) = (f^H)^L(x)$ for all $x \neq 0$ and $(f^H)^L \in Q^L$. This together with the fact that

$$f^{HH}(0) = \inf\{f^{HH}(x) : x \in R^n\}$$

implies $f^{HH} \in Q^L$. By definition,

$$\begin{aligned} f^{HH}(0) &= -\sup\{f^H(y) : y \in R^n\} \\ &= -\sup_{y\in R^n}\{-\inf\{f(x) : \langle y,x\rangle \ge 1\}\} \\ &= \inf_{y\in R^n}\inf_{x}\{f(x) : \langle y,x\rangle \ge 1\} \\ &= \inf_{x}\inf_{y\in R^n}\{f(x) : \langle y,x\rangle \ge 1\} \\ &= \inf\{f(x) : x \in R^n \setminus \{0\}\}. \end{aligned} \tag{31}$$

For $\alpha > \inf\{f(x) : x \in R^n \setminus \{0\}\} = f^{HH}(0)$ one has

$$\begin{aligned} \{x : f^{HH}(x) \le \alpha\} &= \{x : (f^H)^L(x) \le \alpha\} \\ &= \{y : f^H(y) < -\alpha\}^0 \\ &= \{y : f^U(y) < -\alpha\}^0 \\ &= \{\operatorname{int}\{x : f(x) \le \alpha\}^0\}^0. \end{aligned} \tag{32}$$

Since $f \in G^L$, $\{x : f(x) \le \alpha\}$ is either a compact set or the hull space. If $\{x : f(x) \le \alpha\}$ is compact, from (32) it follows that

$$\begin{aligned} \{x : f^{HH}(x) \le \alpha\} &= \{\operatorname{int}\{x : f(x) \le \alpha\}^0\}^0 \\ &= \{x : f(x) \le \alpha\}^{00} \\ &= \operatorname{conv}(\{x : f(x) \le \alpha\} \cup \{0\}). \end{aligned}$$

If $\{x : f(x) \le \alpha\} = R^n$, from (32) it follows that

$$\begin{aligned}\{x : f^{HH}(x) \le \alpha\} &= \{\operatorname{int}\{x : f(x) \le \alpha\}^0\}^0 \\ &= \{\emptyset\}^0 = R^n \\ &= \operatorname{conv}(\{x : f(x) \le \alpha\} \cup \{0\}). \qquad \square\end{aligned}$$

4. Dual problems. Denote by Q the set of functions f such that $f^{HH} = f$. Since f^{HH} is quasi-convex and $f^{HH}(0) = \inf\{f^{HH}(x) : x \neq 0\}$ [19], if $f \in Q$ then f is quasi-convex and

$$(33) \qquad f(0) = \inf\{f(x) : x \neq 0\}.$$

Denote by G the class of quasi-convex functions satisfying (33). Obviously, $Q \subseteq G$. However the difference between Q and G is not very big, because $f \in G$ will be in Q if f is either usc or lsc [19], [22]. For $f \in Q$ we define

$$\begin{aligned}\operatorname{dom} f &= \{x : f(x) < \sup_{x \in R^n} f(x)\}, \\ \ker f &= \{x : f(x) = f(0)\}.\end{aligned}$$

From (33) it follows that $\ker f = \{x : f(x) \le f(0)\}$. Therefore, $\operatorname{dom} f$ and $\ker f$ are convex sets. If f is nonconstant and convex, then $\operatorname{dom} f$ has the usual meaning:

$$\operatorname{dom} f = \{x : f(x) < +\infty\},$$

because $\sup\{f(x) : x \in R^n\} = +\infty$ [14].

Consider the following primal problem:

$$(34) \qquad \sup\{f(x) : x \in D\},$$

where $f \in Q$ and D is a compact set. We call the dual problem of problem (34) the following problem:

$$(35) \qquad \inf\{f^H(y) : y \notin \operatorname{int} D^0\}.$$

THEOREM 4.1. (i) $\sup(34) = -\inf(35)$.

(ii) *If the primal* (34) *is regular then the dual* (35) *is regular.*

(iii) *If y is an optimal solution to the dual* (35) *then every vector $x \in D$ such that $\langle y, x\rangle \ge 1$ is optimal to the primal* (34).

Proof. (i) One has

$$\begin{aligned}&\sup\{f(x) : x \in D\} \\ &\quad = \sup\{f^{HH}(x) : x \in D\} \quad (\text{since } f^{HH} = f) \\ &\quad = \sup_{x \in D}\{-\inf_{y}\{f^H(y) : \langle y, x\rangle \ge 1\}\} \\ &\quad = -\inf_{x \in D}\inf_{y}\{f^H(y) : \langle y, x\rangle \ge 1\} \\ &\quad = -\inf_{y}\inf_{x \in D}\{f^H(y) : \langle y, x\rangle \ge 1\} \\ &\quad = -\inf_{y \notin \operatorname{int} D^0} f^H(y) \quad \left(\text{since } y \in \operatorname{int} D^0 \Leftrightarrow \sup_{x \in D}\langle y, x\rangle < 1\right).\end{aligned}$$

(ii) Suppose that the primal (34) is regular, i.e.,

$$\sup\{f(x) : x \in D\} = \sup\{f(x) : x \in \operatorname{int} D\}.$$

Let $\{x_n\}$ be a sequence in $\operatorname{int} D$ such that

$$f(x_n) \to \sup\{f(x) : x \in D\}. \tag{36}$$

Since $f^{HH} = f$, one has

$$f(x_n) = f^{HH}(x_n) = -\inf\{f^H(y) : \langle y, x_n\rangle \geq 1\}.$$

Let y_n be a vector such that $\langle y_n, x_n\rangle \geq 1$ and

$$| f^H(y_n) + f(x_n) |\leq 1/n. \tag{37}$$

Then, $y_n \notin D^0$ because

$$1 = \langle y_n, x_n\rangle < \sup_{x\in D}\langle y_n, x\rangle \quad (\text{since } x_n \in \operatorname{int} D).$$

Furthermore, from (36) and (37) it follows that

$$-f^H(y_n) \to \sup_{x\in D}\{f(x) : x \in D\} = -\inf_{y\notin \operatorname{int} D^0} f^H(y) \quad \text{as } n \to \infty.$$

Therefore,

$$\inf\{f^H(y) : y \notin D^0\} = \inf\{f^H(y) : y \notin \operatorname{int} D^0\}.$$

So the dual (35) is regular.

(iii) Let y be an optimal solution to the dual (35). Since $y \notin \operatorname{int} D^0$ and D is compact, there is $x \in D$ such that $\langle y, x\rangle \geq 1$. Then,

$$f(x) \geq \inf\{f(z) : \langle y, z\rangle \geq 1\} = -f^H(y) = -\inf(35) = \sup(34).$$

Therefore, x is optimal to the primal (34). □

If the primal problem (34) is nontrivial, i.e., f is nonconstant on D, then

$$\inf(35) = -\sup(34) < -f(0) = -f^{HH}(0) = \sup\{f^H(y) : y \in R^n\}.$$

Therefore,

$$\inf\{f^H(y) : y \notin \operatorname{int} D^0\} \Leftrightarrow \inf\{f^H(y) : y \notin \operatorname{int} D^0, y \in \operatorname{dom} f^H\}.$$

If f is usc, then

$$\begin{aligned}
\operatorname{dom} f^H &= \cup_{\alpha<\sup_{u\in R^n} f^H(u)}\{v : f^H(v) \leq \alpha\}\\
&= \cup_{\alpha<-f(0)}\{v : f^H(v) \leq \alpha\}\\
&= \cup_{\alpha<f(0)}\{x : f(x) < -\alpha\}^0\\
&= \cup_{\beta>f(0)}\{x : f(x) < \beta\}^0\\
&\subseteq (\ker f)^0 \quad (\text{because } \ker f \subseteq \{x : f(x) < \beta\}\ \forall \beta > f(0)).
\end{aligned}$$

This implies that the dual problem can be embedded into a k-dimensional space if $\ker f$ contains an $(n-k)$-dimensional subspace.

Example 4.1 (Burkard, Oettli, and Thach [1]). In a three-dimensional space we are given n vectors $v_i = (a_i, b_i, c_i) \in R^3$, $i = 1, \ldots, n$. Each vector v_i is associated with a weight w_i. We wish to find vectors $v_{i_1}, \ldots, v_{i_k}$ such that the sum of their weights $w_{i_1} + \cdots + w_{i_k}$ is less than or equal to 1, and the vector $v = v_{i_1} + \cdots + v_{i_k}$ has maximum length. This is a generalized knapsack problem. Let $a = (a_1, \ldots, a_n)$, $b = (b_1, \ldots, b_n)$, $c = (c_1, \ldots, c_n)$. If there is a vector $z \in R^n$ such that $a = \alpha z$; $b = \beta z$; $c = \gamma z$, then $a + b + c = (\alpha + \beta + \gamma)z$ and this problem is nothing but the one-dimensional knapsack problem. So, we assume that a, b, c are linearly independent. By using Boolean variables $x_i \in \{0, 1\}$ $i = 1, \ldots, n$ we can formulate the problem as

$$\max\{f(x) : x \in D\},$$

with

$$f(x) := \left(\sum_{i=1}^{n} a_i x_i\right)^2 + \left(\sum_{i=1}^{n} b_1 x_i\right)^2 + \left(\sum_{i=1}^{n} c_i x_i\right)^2$$

$$D := \left\{x = (x_1, ..., x_n) \in \{0, 1\}^n : \sum_{i=1}^{n} w_i x_i \le 1\right\}.$$

Since ker$f$ is the following $(n - 3)$-dimensional subspace

$$C := \left\{x : \sum_{i=1}^{n} a_i x_i = 0, \sum_{i=1}^{n} b_i x_i = 0, \sum_{i=1}^{n} c_i x_i = 0\right\},$$

its dual problem can be embedded into a three-dimensional space as follows:

$$\min\left\{f^H(y) : y \in C^\perp, y \notin \operatorname{int}(D^0)\right\},$$

where

$$C^\perp = \{\alpha a + \beta b + \gamma c : (\alpha, \beta, \gamma) \in R^3\},$$

$$f^H(\alpha a + \beta b + \gamma c) = -\inf\left\{f(x) : \sum_{i=1}^{n} x_i(\alpha a_i + \beta b_i + \gamma c_i) \ge 1\right\}.$$

Since $f(\cdot)$ is convex quadratic, the value $f^H(\alpha a + \beta b + \gamma c)$ can be immediately computed:

$$f^H(\alpha a + \beta b + \gamma c) = -\frac{1}{\alpha^2 + \beta^2 + \gamma^2}.$$

Furthermore, the constraint $(\alpha a + \beta b + \gamma c) \notin \operatorname{int}(D^0)$ is equivalent to

$$\max\left\{\sum_{i=1}^{n} x_i(\alpha a_i + \beta b_i + \gamma c_i) : x \in \{0, 1\}^n\right\} \ge 1.$$

Hence the dual problem can be rewritten as

$$\min\left\{-\frac{1}{\alpha^2 + \beta^2 + \gamma^2} : \max\left\{\sum_{i=1}^{n} x_i(\alpha a_i + \beta b_i + \gamma c_i) : x \in \{0, 1\}^n\right\} \ge 1\right\}$$

$$\Leftrightarrow \min\left\{\alpha^2 + \beta^2 + \gamma^2 : \max\left\{\sum_{i=1}^{n} x_i(\alpha a_i + \beta b_i + \gamma c_i) : x \in \{0, 1\}^n\right\} \ge 1\right\},$$

which is a convex minimization over the complement of a convex set in R^3.

We present now a dual problem for a general minimization over the complement of a convex set:

$$\inf\{f(x) : x \notin \operatorname{int} D\}, \tag{38}$$

where f is an arbitrary function and D is a closed convex set such that $0 \in \operatorname{int} D$. It is well known that $D^{00} = D$. We call the dual problem of problem (38) the following problem:

$$\sup\{f^H(y) : y \in D^0\}. \tag{39}$$

THEOREM 4.2. (i) $\inf(38) = -\sup(39)$;

(ii) *If the primal* (38) *is regular and D is bounded then the dual* (39) *is regular*;

(iii) *If y is an optimal solution to the dual* (39) *then every minimizer of the function f on the half-space $\{x : \langle y, x\rangle \geq 1\}$ is an optimal solution to the primal* (38).

Proof. (i) One has

$$\begin{aligned}
&-\sup\{f^H(y) : y \in D^0\}\\
&\quad = -\sup_{y\in D^0}\{-\inf_x\{f(x) : \langle y, x\rangle \leq 1\}\}\\
&\quad = \inf_{y\in D^0}\inf_x\{f(x) : \langle y, x\rangle \geq 1\}\\
&\quad = \inf_x \inf_{y\in D^0}\{f(x) : \langle y, x\rangle \geq 1\}\\
&\quad = \inf_{x\notin \operatorname{int} D} f(x) \quad \left(\text{since } x \in \operatorname{int} D = \operatorname{int} D^{00} \Leftrightarrow \sup_{y\in D^0}\langle y, x\rangle < 1\right).
\end{aligned}$$

(ii) Suppose that D is bounded and the primal problem (38) is regular, i.e.,

$$\inf\{f(x) : x \notin \operatorname{int} D\} = \inf\{f(x) : x \notin D\}.$$

Let $\{x_n\}$ be a sequence such that $x_n \notin D$ and

$$f(x_n) \to \inf(38) \quad \text{as} \quad n \to \infty. \tag{40}$$

Let $\langle y_n, \cdot\rangle$ be a linear function such that

$$1 = \langle y_n, x_n\rangle > \sup_{x\in D}\langle y_n, x\rangle. \tag{41}$$

Since D is bounded, y_n belongs to $\operatorname{int} D^0$. Furthermore,

$$\sup(39) \geq f^H(y_n) = -\inf\{f(x) : \langle y_n, x\rangle \geq 1\} \geq -f(x_n).$$

This together with (40) and assertion (i) implies

$$f^H(y_n) \to \sup(39).$$

Since $y_n \in \operatorname{int} D^0$, this means that the dual (39) is regular.

(iii) Let y be an optimal solution to the dual (39) and z be a minimizer of f on the half-space $\{x : \langle y, x\rangle \geq 1\}$. Then,

$$\begin{aligned}
f(z) &= \inf\{f(x) : \langle y, x\rangle \geq 1\} = -f(y)\\
&= -\sup(39) = \inf(38).
\end{aligned}$$

Since $y \in D^0$ and $\langle y, z\rangle \geq 1$, z does not belongs to $\operatorname{int} D$. Therefore z is optimal to the primal (38). $\square$

Example 4.2 (Burkard, Oettli, and Thach [1]). As in Example 4.1 we are given n vectors $T_i = (a_i, b_i, c_i) \in R^3$ $i = 1, \ldots, n$. Each vector T_i is associated with an weight w_i. This time we wish to find vector $T_{i_1}, T_{i_2}, \ldots, T_{i_k}$ such that the length of the vector $T = T_{i_1} + \cdots + T_{i_k}$ is greater than or equal to 1 and the sum of their weights $w_{i_1} + \cdots + w_{i_k}$ is minimized. By using a Boolean variable $x \in \{0,1\}^n$ we formulate the problem as follows:

$$\min\{\langle w, x\rangle : \langle a, x\rangle^2 + \langle b, x\rangle^2 + \langle c, x\rangle^2 \geq 1, x \in \{0,1\}^n\}, \tag{42}$$

where $w = (w_1, \ldots, w_n), a = (a_1, \ldots, a_n), b = (b_1, \ldots, b_n)$, and $c = (c_1, \ldots, c_n)$. By setting

$$f(x) = \begin{cases} \langle w, x\rangle & \text{if } x \in \{0,1\}^n \\ +\infty & \text{otherwise,} \end{cases}$$

$$D \;= \{x : \langle a, x\rangle^2 + \langle b, x\rangle^2 + \langle c, x\rangle^2 \leq 1\},$$

the problem is rewritten as follows:

$$\inf\{f(x) : x \notin \operatorname{int} D\}.$$

Then the dual is

$$\sup\{f^H(y) : y \in D^0\}.$$

Since $\{x : \langle a, x\rangle = 0, \langle b, x\rangle = 0, \langle c, x\rangle = 0\} \subseteq D$, D^0 can be embedded into the three-dimensional space:

$$D^0 = \{\alpha.a + \beta.b + \gamma.c : \alpha^2 + \beta^2 + \gamma^2 \leq 1\}.$$

Therefore the dual is a quasi-convex maximization in the three-dimensional space:

$$\sup\{f^H(\alpha.a + \beta.b + \gamma.c) : \alpha^2 + \beta^2 + \gamma^2 \leq 1\}.$$

Note that for each (α, β, γ) we can compute

$$f^H(\alpha.a + \beta.b + \gamma.c) = -\min\{\langle w, x\rangle : \langle \alpha.a + \beta.b + \gamma.c, x\rangle \geq 1, x \in \{0,1\}^n\}$$

by a pseudopolynomial dynamic programming algorithm.

As presented above, a dual of a quasi-convex maximization over a compact set is a quasi-convex minimization over the complement of a convex set and a dual of a general minimization over the complement of a convex set is a quasi-convex maximization over a compact convex set. In the remainder of this section we discuss the duality in another class of nonconvex programs:

$$\inf_{x \in R^n} \{h_1(x) - h_2(x)\}, \tag{43}$$

where h_1 is an arbitrary function and h_2 is a finite convex function. In Toland [25], and Hiriart-Urruty [4] a dual of problem (43) is

$$\inf_{y \in \operatorname{dom} h_2^*} \{h_2^*(y) - h_1^*(y)\}. \tag{44}$$

In the sequel we shall see that if problem (43) is regarded as a particular case of problem (38), then the dual (44) can be obtained by our duality scheme. By adding a variable t we transform problem (43) into the following problem:

$$\inf\{h_1(x) - t : h_2(x) - t - h_2(0) - 1 \geq 0\}. \tag{45}$$

By setting

$$f(x,t) = h_1(x) - t,$$
$$D = \{(x,t) : h_2(x) - t - \alpha \leq 0\},$$

where $\alpha = h_2(0) + 1$, problem (45) has the form (38). Note that $0 \in \operatorname{int} D$, since $\alpha = h_2(0) + 1$.

PROPOSITION 4.3. *Let h be an arbitrary function on R^n. Set $g(x,t) = h(x) - t$, $Y = \{y \in R^n : \langle y, x\rangle < 1$ for all $x \in \operatorname{dom} h\}$. Then,*

$$g^H(y,\lambda) = \begin{cases} (1/\lambda) + h^*(-y/\lambda) & \text{if } \lambda < 0 \\ -\infty & \text{if } y \in Y \text{ and } \lambda = 0 \\ +\infty & \text{otherwise.} \end{cases}$$

Proof. By definition of the quasi conjugate, it is easy to see that $g^H(y) = -\infty$ if $y \in Y$ and $\lambda = 0$, and $g^H(y,\lambda) = +\infty$ if either $\lambda > 0$ or $y \notin Y$ and $\lambda = 0$. If $\lambda < 0$ then

$$\begin{aligned} g^H(y,\lambda) &= -\inf\{h(x) - t : \langle y,x\rangle + \lambda.t \geq 1\} \\ &= -\inf\left\{h(x) - t : -t \geq \left\langle \frac{y}{\lambda}, x\right\rangle - \frac{1}{\lambda}\right\} \\ &= -\inf_{x\in R^n}\left\{h(x) + \left\langle \frac{y}{\lambda}, x\right\rangle - \frac{1}{\lambda}\right\} \\ &= \sup\left\{\left\langle -\frac{y}{\lambda}, x\right\rangle - h(x) + \frac{1}{\lambda}\right\} \\ &= h^*\left(-\frac{y}{\lambda}\right) + \frac{1}{\lambda}. \end{aligned}$$

This completes the proof. □

Setting $Y_1 = \{y \in R^n : \langle y,x\rangle < 1 \text{ for all } x \in \operatorname{dom} h_1\}$, by Propositions 2.3 and 4.3, one has

$$f^H(y,\lambda) = \begin{cases} 1/\lambda + h_1^*(-\frac{y}{\lambda}) & \text{if } \lambda < 0, \\ -\infty & \text{if } y \in Y_1 \text{ and } \lambda = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

$$D^0 = \left\{(y,\lambda) : \lambda < 0, h_2^*\left(-\frac{y}{\lambda}\right) + \alpha \leq -\frac{1}{\lambda}\right\} \cup \{0\}.$$

The dual of problem (45) is maximizing f^H on D^0. Since $f^H(0,0) = -\infty$, maximizing f^H on D^0 is equivalent to maximizing f^H on $D^0 \setminus \{0\}$:

$$\begin{aligned} &\sup\left\{h_1^*\left(-\frac{y}{\lambda}\right) + \frac{1}{\lambda} : \lambda < 0, h_2^*\left(-\frac{y}{\lambda}\right) + \alpha \leq -\frac{1}{\lambda}\right\} \\ &= \sup\left\{h_1^*\left(-\frac{y}{\lambda}\right) - h_2^*\left(-\frac{y}{\lambda}\right) - \alpha : \lambda < 0, y \in \operatorname{dom} h_2^*\right\} \\ &= \sup\{h_1^*(y) - h_2^*(y) : y \in \operatorname{dom} h_2^*\} - \alpha \\ &= -\inf\{h_2^*(y) - h_1^*(y) : y \in \operatorname{dom} h_2^*\} - \alpha. \end{aligned}$$

Thus, we obtain the dual (44).

5. Applications.

Application 5.1. We are given a convex maximization program

$$\sup\{f(x) : h(x) \leq 1\}, \tag{46}$$

where f and h are the Minkowski functionals of the compact convex sets C and D such that $0 \in C$ and $0 \in D$. It is well known that

$$\begin{aligned} f(x) &= \sup\{\langle y, x\rangle : y \in C^0\}, \\ h(x) &= \sup\{\langle y, x\rangle : y \in D^0\}. \end{aligned}$$

Consider a perturbation of program (46):

$$(A(z)) \qquad \sup\{f(x) : h(x) \leq 1 + \langle z, x\rangle\}.$$

It is obvious that program $(A(z))$ is program (46) when $z = 0$. We wish to solve the following parametric problem:

$$s(z) := \sup(A(z)) \rightarrow \inf, \qquad \text{such that } z \in R^n. \tag{47}$$

If $z \notin \operatorname{int} D^0$ then there is vector x such that

$$1 = \langle z, x\rangle \geq \sup\{\langle y, x\rangle : y \in D^0\} = h(x) > 0.$$

So,

$$h(\lambda.x) = \lambda.h(x) \leq \lambda.\langle z, x\rangle < \langle z, \lambda.x\rangle + 1 \quad \forall\lambda \geq 0.$$

Therefore, the half line $\{\lambda.x : \lambda \geq 0\}$ is contained in the feasible domain of program $(A(z))$. This implies $s(z) = +\infty$. If $z \in \operatorname{int} D^0$, then $0 \in \operatorname{int}(D^0 - z)$. Since

$$\begin{aligned} (D^0 - z)^0 &= \left\{x : \sup_{y \in D^0} \langle y - z, x\rangle \leq 1\right\} \\ &= \{x : h(x) \leq 1 + \langle z, x\rangle\}, \end{aligned}$$

the dual of $(A(z))$ is

$$\inf\{f^H(y) : y \notin \operatorname{int}(D^0 - z)\}. \tag{48}$$

It is easy to check that

$$f^H(y) = -1/\sup\{\langle y, x\rangle : x \in C\}.$$

The dual (48) is equivalent to

$$\inf\{g(y) : y \notin \operatorname{int}(D^0 - z)\},$$

where $g(y) = \sup\{\langle y, x\rangle : x \in C\}$, or equivalently,

$$\inf\{g(y - z) : y \notin \operatorname{int} D^0\}. \tag{49}$$

Set $r(z) = \inf(49)$. By the duality, $s(z) = -\inf(48)$, and hence $s(z) = 1/\inf(49) = 1/r(z)$. Therefore, the parametric problem (47) is equivalent to

$$(50) \qquad r(z) \to \sup \quad \text{such that } z \in \operatorname{int} D^0.$$

Since $g(.)$ is the Minkowski functional of the convex set C^0, $r(z)$ is a concave function on D^0 [18] and hence problem (50) is a concave maximization over a convex set, i.e., a convex program. If $f(x) =\| x \|$ then $g(y) =\| y \|$ and program (50) is used to find the biggest ball contained in D^0.

Application 5.2. We are given a convex minimization over the complement of a convex set:

$$(51) \qquad \inf\{f(x) : h(x) \geq 1\},$$

where f and h are functions as in Application 5.1. We assume additionally that $0 \in \operatorname{int} D$. Consider a perturbation of program (51):

$$(B(z)) \qquad \inf\{f(x) : h(x) \geq 1 + \langle z, x\rangle\},$$

and we wish to solve the following parametric problem

$$(52) \qquad s(z) := \inf(B(z)) \to \sup \quad \text{such that } z \in R^n.$$

For every $z \in R^n$, one has

$$\begin{aligned}(D^0 - z)^0 &= \left\{x : \sup_{y \in D^0} \langle y - z, x\rangle \leq 1\right\} \\ &= \{x : h(x) \leq 1 + \langle z, x\rangle\}\end{aligned}$$

and hence

$$\{x : h(x) \leq 1 + \langle z, x\rangle\}^0 = (D^0 - z)^{00} = \operatorname{conv}\{(D^0 - z) \cup \{0\}\}.$$

The dual of (B(z)) is

$$\begin{aligned}(53) \qquad & \sup\{f^H(y) : y \in \operatorname{conv}\{(D^0 - z) \cup \{0\}\}\} \\ & \Leftrightarrow \sup\{f^H(y) : y \in D^0 - z\} \\ & \Leftrightarrow \sup\{f^H(y - z) : y \in D^0\} \\ (54) \qquad & \Leftrightarrow \sup\{g(y - z) : y \in D^0\},\end{aligned}$$

where $g(y) = \sup\{\langle y, x\rangle : x \in C\}$. Set $r(z) = \sup(54)$. By the duality, $s(z) = -\sup(53)$ and hence $s(z) = 1/\sup(54) = 1/r(z)$ because $f^H(y) = -1/g(y)$. Therefore, the parametric problem (52) is equivalent to

$$(55) \qquad r(z) \to \inf \quad \text{such that } z \in R^n.$$

Since $g(.)$ is convex, $r(.)$ is convex. So problem (55) is a convex program. If $f(x) = \| x \|$ then $g(y) =\| y \|$ and problem (55) is used to find the smallest ball containing D^0.

Application 5.3 (Konno, Thach, and Yokota [7]). We consider a multi-objective problem:

$$\langle c_i, x\rangle \to \max \ i = 1, \ldots, p \quad \text{such that } x \in X,$$

where $\langle c_i, \cdot \rangle, i = 1, \dots, p$ are p objectives and X is a polyhedral convex set of feasible solutions. We denote by X^* the set of nondominated solutions and we wish to find a solution in X^* which minimizes an additional objective function $f : R^n \mapsto R$:

$$(56) \qquad f(x) \to \min \quad \text{such that } x \in X^*.$$

For each $s > 0$ we define

$$\begin{aligned} c_i^s &= c_i + \frac{1}{s} \sum_{i=1}^{p} c_i, \\ C^s &= \{x \in R^n : \langle c_i, x \rangle \le 0 \; i = 1, \dots, p\}, \\ X^s &= \cup_{(t_1, \dots, t_p) \ge 0, \sum_{i=1}^{p} t_i = 1} \operatorname{argmax} \left\{ \sum_{i=1}^{p} t_i c_i^s : x \in X \right\}. \end{aligned}$$

Then X^s can be represented as follows:

$$X^s = X \setminus \operatorname{int}\{X + C^s\}$$

and $X^* = X^s$ for sufficiently large enough s (Yu [29], Sawaragi, Nakayama, and Tanino [15]). Therefore, problem (56) is equivalent to

$$(57) \qquad f(x) \to \min \quad \text{such that } x \in X \setminus \operatorname{int}\{X + C^s\}.$$

The constraint in (57) is a reverse convex constraint. By translating the original into $\operatorname{int}\{X + C^s\}$, if necessary, we can assume that $0 \in \operatorname{int}\{X + C^s\}$. The dual of (57) is

$$(58) \qquad f^H(y) \to \max \quad \text{such that } y \in (X + C^s)^0.$$

Since C^s contains the subspace $\{x : \langle c_i^s, x \rangle = 0 \text{ for all } i = 1, \dots, p\}$, the dual (58) can be embedded into the p-dimensional space:

$$(59) \; F(t_1, \dots, t_p) \to \max$$

$$\text{such that} \quad \sup \left\{ \left\langle \sum_{i=1}^{p} t_i c_i^s, x \right\rangle : x \in X \right\} \le 1, \quad \sum_{i=1}^{p} t_i = 1, \quad t_1 \ge 0, \dots, t_p \ge 0,$$

where

$$(60) \qquad F(t_1, \dots, t_p) = -\inf \left\{ f(x) : \left\langle \sum_{i=1}^{p} t_i c_i, x \right\rangle \ge 1, x \in X \right\}.$$

If f is convex, then (60) is a convex program. The dual (59) is a quasi-convex maximization over a compact convex set in R^p. In practice the number of objective functions, p, is often very small compared with the number of variables, n. Therefore the dual problem is much easier to handle than the primal by the existing solution methods (Horst and Tuy [6]).

6. Discussions. Let us discuss on the following classes of multi-extremal problems:

(QMIN) Quasi-convex MINimization over a convex set;
(QMINR) Quasi-convex MINimization over the complement of a convex set;
(QMAX) Quasi-convex MAXimization over a convex set.

Since these classes contain convex programs, it would be interesting if we could find connections between optimality criteria, duality schemes in these classes and convex programs. There have been many significant results for extending the conjugation and convex duality to QMIN problems (see, e.g., Greenberg and Pierskalla [3], Crouzeix [2], Oettli [10], [11], Tind and Wolsey [26], Passy and Prisman [12], Singer [6], [7], Martinez-Legaz [8], Penot and Volle [13]). For a survey we refer to Penot and Volle [13]. Since the class of quasi-convex functions is much bigger than that of convex functions, we have to use an extra parameter in the extensions. More recently, in [23], we have proved that if we restrict quasi-convex functions to a reasonable class, then we can obtain generalizations, which involve no extra parameter. In [24] we present a generalized Kuhn–Tucker condition and a variational inequality for QMIN, and show that the path-following methods could, in principle, be applied to QMIN. Thus QMIN problems are multi-extremal, but they are reasonably closer to convex problems.

A duality scheme for QMINR and QMAX was presented in [19]. We have proved that this duality could be obtained from convex duality and the minimax principles, and if we consider a convex problem as a particular case of convex maximization problems under convex constraints; then the dual problems by convex duality and by the generalized duality are essentially equivalent [22]. In this paper we show a connection between the nonconvex duality and geometrical optimality criteria by using the polar relationships of functions and their quasi conjugates. As we have seen, the optimality criteria and duality in QMINR and QMAX are quite different from those in convex programs. By conjugations we cannot reduce QMINR or QMAX problems to solving equations (or generalized equations). This might be a reason that explains why the algorithms for QMINR and QMAX problems have so far been constructed on the basis of enumeration and branch-and-bound methods, which are completely different from path-following methods and whose complexity is exponential with resepct to the problem dimension. However, we can use the nonconvex duality to reduce the dimension of certain large-scale nonconvex problems and hence to obtain practical algorithms for solving them [1], [7], [20], [21], [27], [28].

Acknowledgments. The author expresses his thanks to Professor W. Oettli for helpful remarks on the relations between the optimality criterion and the duality in optimization problems. The author thanks the referees for many valuable comments and suggestions.

REFERENCES

[1] R. E. Burkard, W. Oettli and P. T. Thach, *Dual solution methods for two discrete optimization problems in the space*, Institute of Mathematics, Graz University of Technology, Graz, Austria, 1991, preprint.

[2] J.-P. Crouzeix, *A duality framework in quasiconvex programming*, in Generalized Concavity in Optimization and Economics, S. Schaible and W. T. Ziemba, eds., Academic Press, New York, 1981, pp. 207–225.

[3] H. P. Greenberg and W. P. Pierskalla, *Quasiconjugate functions and surrogate duality*, Cahiers Centre Etudes Rech. Oper. 15(1973), pp. 437–448.

[4] J.-B. Hiriart-Urruty, *Generalized differentiability, duality and optimization for problems dealing with the difference of convex functions*, Lecture Notes in Econom. and Math. Systems, Springer-Verlag, Berlin, 256 (1985), pp. 37–69.

[5] ———, *Conditions necessaires et suffisantes d'optimalite globale en optimisation de differences de deux fonctions convexes*, Compte Rendus de l'Academie des Sciences, Paris, Serie 1(1989), pp. 459–462.

[6] R. Horst and H. Tuy, *Global Optimization*, Springer-Verlag, New York, 1990.

[7] H. Konno, P. T. Thach, and Y. Yokota, *A dual approach to a minimization on the set of Pareto-optimal solutions*, 1992, manuscript.

[8] J. E. Martinez-Legaz, *Quasiconvex duality theory by generalized conjugation methods*, Optimization 19(1988), pp. 603–652.

[9] G. L. Nemhauser, A. H. G. Rinnooy Kan, and M. J. Todd, *Handbooks in Operations Research and Management Science*, Vol. 1: Optimization, North-Holland, Amsterdam, 1989.

[10] W. Oettli, *Optimality conditions involving generalized convex mappings*, Generalized Concavity in Optimization and Economics, S. Schaible and W. T. Ziemba, eds., Academic Press, New York, 1981, pp. 227–238.

[11] ———, *Optimality conditions for programming problems involving multivalued mapping*, in Applied Mathematics, B. Korte, ed., North-Holland, Amsterdam, 1982, pp. 196–226.

[12] U. Passy and E. Z. Prisman, *A convex-like duality scheme for quasiconvex programs*, Mathematical Programming 32(1985), pp. 278–300.

[13] J. P. Penot and M. Volle, *On quasi-convex duality*, Math. Oper. Res., 15(1990), pp. 597–625.

[14] R. T. Rockafellar, *Convex analysis*, Princeton University Press, Princeton, NJ, 1970.

[15] Y. Sawaragi, H. Nakayama and T. Tanino, *Theory of multiobjective optimization*, Academic Press Inc. LTD., London, 1985.

[16] I. Singer, *Optimization by level set methods,* VI: *Generalization of surrogate type reverse convex duality*, Optimization 18(1987), pp. 485–499.

[17] ———, *A general theory of dual optimization problems*, J. Math. Anal. Appl., 116 (1986), pp. 77–130.

[18] P. T. Thach, *The design centering problem as a d.c. programming problem*, Math. Programming, 41 (1988), pp. 229–248.

[19] ———, *Quasiconjugates of functions, duality relationship between quasiconvex minimization under a reverse convex constraint and quasiconvex maximization under a reverse convex constraint and applications*, J. Math. Anal. Appl., 159 (1991), pp. 299–322.

[20] P. T. Thach, R. E. Burkard and W. Oettli, *Mathematical programs with a two-dimensional reverse convex constraint*, J. Global Optimization, 1(1991), pp. 145–154.

[21] P. T. Thach and H. Tuy, *Dual outer approximation methods for concave programs and reverse convex programs*, Report N.90-30, Institute of Human and Social Sciences, Tokyo Institute of Technology, 1990; Math. Oper. Res., submitted.

[22] P. T. Thach, *A generalized duality and applications*, Report N.90-31, Institute of Human and Social Sciences, Tokyo Institute of Technology, 1990; J. Global Optimization, submitted.

[23] ———, *A nonconvex duality with zero gap and applications*, preprint, Department of Mathematics, Trier University, 1991; SIAM J. Optim., to appear.

[24] P. T. Thach and M. Kojima, *A generalized convexity and variational inequalities for quasiconvex minimization*, Report IHSS 92-52, Tokyo Institute of Technology, 1992.

[25] J. F. Toland, *Duality in nonconvex optimization*, J. Math. Anal. Appl., 66 (1978), pp. 399–415.

[26] J. Tind and L. A. Wolsey, *An elementary survey of general duality theory in mathematical programming*, Math. Programming, 21 (1981), pp. 241–261.

[27] H. Tuy, *Polyhedral annexation, dualization and dimension reduction technique in global optimization*, J. Global Optimization, 1 (1991), pp. 229–244.

[28] H. Tuy, A. Migdalas and P. Warbrand, *A global optimization approach for the linear two-level program*, Report n.90-17, Department of Mathematics, Linkoping University, 1990.

[29] P. L. Yu, *Multiple Criteria Decision Making: Concepts, Techniques, and Extensions*, Plenum Press, New York, 1985.

SIAM J. MATH. ANAL.
Vol. 24, No. 6, pp. 1557–1576, November 1993

008

MAXIMUM ENTROPY REGULARIZATION FOR FREDHOLM INTEGRAL EQUATIONS OF THE FIRST KIND*

P. P. B. EGGERMONT†

Abstract. The regularization of Fredholm integral equations of the first kind is considered with positive solutions by means of maximum entropy. The regularized solution is the minimizer of a functional analogous to the case of Phillips–Tikhonov regularization. The regularized solution is shown to converge to the solution of the maximum entropy least squares problem, assuming it exists. Under additional regularity conditions akin to those for Phillips–Tikhonov regularization error estimates are obtained as well. In addition it is shown that the regularity conditions are necessary for these estimates to hold. Approximation from finite-dimensional subspaces are also considered, as well as exact and approximate moment problems for the integral equations. The basic tools in the analysis are the weak compactness of subsets of L^1 consisting of functions of bounded entropy, and an inequality for convex optimization problems with Bregman functionals.

Key words. Fredholm integral equation of the first kind, ill-posed problem, regularization, maximum entropy, Bregman functional

AMS subject classifications. 45L05, 45B05, 65R20

1. Introduction.

We consider the regularization of ill-posed problems by the maximum entropy method. We are interested in stability and error estimates, and in the convergence of the method as the error in the data tends to zero. There are two basic versions of the maximum entropy method. In the older version the entropy of the unknown density distribution is maximized subject to the equality constraints imposed by the data, while the newer version closely resembles the classical Phillips–Tikhonov regularization of ill-posed least squares problems. Our aim is to get estimates similar to the Philips–Tikhonov theory with positivity constraints; see Butler, Reeds, and Dawson [4] and Groetsch [12]. The sensible modification of the older version in which the data are interpreted as providing inequality constraints was considered by Amato and Hughes [1]. This is also the usual interpretation in the statistical treatment; see Skilling [22]. For the maximum entropy method applied to some classical moment problems Borwein and Lewis [2] have obtained strong results. The maximum entropy method has a rich history: we mention Larkin [17], Gordon and Herman [13], Frieden [10], and quite recently Smith and Zoltani [23] and Amato and Hughes [1]. For a guide to the lively recent literature on the statistical aspects of maximum entropy, see the conference proceedings [22]. Here we do not consider such statistical aspects as these.

As the prototypical example of an ill-posed problem we consider the Fredholm integral equation of the first kind. So let $\Omega, \Sigma \subset I\!R^N$ be bounded closed domains, and let $k \in C(\Sigma \times \Omega)$. Define the (compact) operator $\mathcal{K} : L^1(\Omega) \longrightarrow L^2(\Sigma)$ by

$$\mathcal{K}f(x) = \int_\Omega k(x,y)f(y)\,d\mu(y), \quad x \in \Sigma, \tag{1.1}$$

*Received by the editors May 18, 1992; accepted for publication (in revised form) February 17, 1993.

†Department of Mathematical Sciences, University of Delaware, Newark, Delaware 19716 (eggermon@brahms.udel.edu).

where $d\mu(y)$ denotes Lebesgue measure on $I\!R^N$. We note that the operator $\mathcal{K}$ is also compact as mapping from $L^1(\Omega)$ into $C(\Sigma)$. We consider the equation

$$\mathcal{K}f(x) = g(x), \quad x \in \Sigma, \tag{1.2}$$

with given $g \in L^2(\Sigma)$, possibly subject to error. In general this is an ill-posed problem; see Groetsch [12]: solutions may not exist, or the solution f may not depend continuously on the data g. In typical applications, the solution f is nonnegative, e.g., because it represents a density or intensity. Formulating the problem as a constrained least squares problem, viz.

$$\begin{aligned} &\text{mimimize } \|\mathcal{K}f - g\|^2_{L^2(\Sigma)} \\ &\text{subject to } f \geq 0, \end{aligned} \tag{1.3}$$

does not essentially change the difficulty, though. In this paper we study variations of Phillips–Tikhonov regularization of (1.2) or (1.3), i.e., we consider

$$\begin{aligned} &\text{mimimize} \quad \|\mathcal{K}f - g\|^2_{L^2(\Sigma)} + \alpha^2 D(f, \varphi) \\ &\text{subject to} \quad f \geq 0, \end{aligned} \tag{1.4}$$

where α is a regularization parameter, φ is the choice for f in the absence of any data, and $D(f, \varphi)$ is the smoothing term. The case

$$D(f, \varphi) = \|\mathcal{L}(f - \varphi)\|^2_{L^2(\Omega)}, \tag{1.5}$$

with $\mathcal{L}$ a differential operator, e.g., the Laplacian, corresponds to classical Phillips–Tikhonov regularization; see Groetsch [12]. Here we are interested in the *cross-entropy* functional

$$D(f, \varphi) = \int_\Omega \left\{ f(y) \log \frac{f(y)}{\varphi(y)} + \varphi(y) - f(y) \right\} d\mu(y), \tag{1.6}$$

which is defined for all nonnegative f and φ (but its value could be $+\infty$). Note that the integrand in (1.6) is nonnegative. To get a feel for the relation between (1.5) and (1.6) we quote the inequalities (see Kemperman [15] and also [2])

$$\|f - \varphi\|^2_{L^1(\Omega)} \leq (\tfrac{4}{3}F + \tfrac{2}{3}\Phi)\, D(f, \varphi), \tag{1.7}$$

where $F = \int_\Omega f$, $\Phi = \int_\Omega \varphi$, and

$$D(f, \varphi) \leq \int_{\operatorname{supp} \varphi} \frac{|f(y) - \varphi(y)|^2}{\varphi(y)}\, d\mu(y), \tag{1.8}$$

where we assume that f vanishes almost everywhere where φ vanishes. This last inequality arises from the usual convexity inequality for $D(f, \varphi) - D(f, f) \leq \langle D_\varphi(f, \varphi), \varphi - f \rangle$.

Volumes have been and are being written to justify the scheme (1.4)–(1.6) from a statistical point of view; see [22]. In this paper we are concerned with justifying it as a regularization method, as follows. Suppose that $f \geq 0$ and g satisfy $\mathcal{K}f = g$ and that we are given some $\delta \geq 0$ and g_δ with $\|g_\delta - g\| \approx \delta$. Let $f_{\alpha\delta}$ denote the solution of (1.4) with g replaced by g_δ, i.e., $f_{\alpha\delta}$ solves

$$\begin{aligned} &\text{mimimize} \quad \|\mathcal{K}f - g_\delta\|^2_{L^2(\Sigma)} + \alpha^2 D(f, \varphi) \\ &\text{subject to} \quad f \geq 0. \end{aligned} \tag{1.9}$$

Can we get estimates for $\|f_{\alpha\delta} - f\|$, and does the estimate tend to zero as α, $\delta \longrightarrow 0$ (in some way)? And what is the appropriate norm to use here? In the same vein we may ask about the dependence of $f_{\alpha\delta}$ on g_δ and φ for fixed $\alpha > 0$. Partial qualitative answers to this were obtained by Klaus and Smith [16], *complete* quantitative answers were obtained under very strong technical conditions by Engl and Landl [9]. Essentially the same answers to the above questions but without any such technical conditions derive from the following inequalities for (1.4). If f_α solves (1.4), then for all nonnegative f,

$$\|\mathcal{K}(f - f_\alpha)\|^2_{L^2(\Sigma)} + \alpha^2 D(f, f_\alpha) \leq \ell(f) - \ell(f_\alpha), \tag{1.10}$$

where $\ell(f) = \|\mathcal{K}f - g\|^2_{L^2(\Sigma)} + \alpha^2 D(f, \varphi)$, the objective function in (1.4), and if F_α solves (1.4) with g replaced by G, then

$$\|\mathcal{K}(F_\alpha - f_\alpha)\|^2_{L^2(\Sigma)} + \alpha^2 D(F_\alpha, f_\alpha) \leq 4\,\|G - g\|^2_{L^2(\Sigma)}. \tag{1.11}$$

These inequalities contain all the required information. Note that (1.11) shows a well-posedness aspect of the problem (1.4): the quantity $\mathcal{K}f_\alpha$ is well behaved uniformly in α, whereas for fixed $\alpha > 0$ the solution f_α itself is well behaved. The inequality (1.10) is actually quite well known. Note that $\|Kf - g\|^2_{L^2(\Sigma)}$ and $D(f, \varphi)$ as given by (1.6) may be written as

$$\|\mathcal{K}f - g\|^2_{L^2(\Sigma)} = N(\mathcal{K}f) - N(g) - \langle N'(g), \mathcal{K}f - g\rangle, \tag{1.12}$$

$$D(f, \varphi) = d(f) - d(\varphi) - \langle d'(\varphi), f - \varphi\rangle, \tag{1.13}$$

where $N(f) = \|f\|^2_{L^2(\Sigma)}$, and $d(f) = \int_\Omega f(y) \log f(y)\, d\mu(y)$, and $\langle \cdot\,, \cdot\rangle$ denotes the inner product on $L^2(\Omega)$ and $L^2(\Sigma)$. Here $N'(f)$ and $d'(f)$ are the Gateaux derivatives of $N(f)$ and $d(f)$ with respect to f. The inequality (1.10) is now nothing more than the statement that N and d are *convex* functions; see Bregman [3] and also [5]. Recently Chen and Teboulle [6] have made use of this as well. In the entropy context Csiszår and Tusnàdy [7] have a similar inequality. We reproduce the proof in §3. The inequality (1.11) is a corollary to (1.10). All this suggests that we consider more than just maximum entropy regularization, i.e., consider (1.4) with D any function satisfying (1.13) for some convex d. Besides (1.5) and (1.6) there are also

$$D(f, \varphi) = \int_\Omega \left\{ \log \frac{\varphi(y)}{f(y)} + \frac{f(y)}{\varphi(y)} - 1 \right\} d\mu(y), \tag{1.14}$$

which is known as Burg's entropy (see [24] for some applications) and

$$D(f, \varphi) = \int_\Omega \left\{ \log \frac{\cosh f(y)}{\cosh \varphi(y)} - (f(y) - \varphi(y)) \tanh \varphi(y) \right\} d\mu(y), \tag{1.15}$$

which arises from $d(f) = \int_\Omega \log\cosh(f(y))\,d\mu(y)$. This choice, or rather a discrete approximation to $\Sigma_j D(h\partial_{x_j} f, \mathbf{0}_\Omega)$, with h a stepsize, is suggested by Green [11]. However, meaningful results can be obtained only if an inequality like (1.7) holds.

Maximum entropy regularization of Fredholm integral equations of the first kind has been the subject of some recent studies. Klaus and Smith [16] study a discrete version of (1.9) in the $L^2(\Omega)$-setting, and prove weak convergence. Engl and Landl [9] study (1.9) (with a slightly different cross-entropy functional) by translating (1.9) into a Phillips–Tikhonov regularization problem for a nonlinear equation, and then applying their general theory of such problems. The translated problem of Engl and Landl [9] reads

$$\text{(1.16)}\qquad \begin{aligned} &\text{mimimize} && \|\mathcal{K}T(u) - g_\delta\|^2_{L^2(\Sigma)} + \alpha^2 \|\, u\,\|^2_{L^2(\Omega)} \\ &\text{subject to} && u \geq 0, \end{aligned}$$

where $T : \mathcal{D}(T) \subset L^1(\Omega) \longrightarrow L^2(\Omega)$ is an explicitly determined bijective *nonlinear* operator such that if $f = T(u)$ then

$$\text{(1.17)}\qquad \|\, u\,\|^2_{L^2(\Omega)} = \int_\Omega \left\{ f(y) \log \frac{f(y)}{\varphi(y)} + \frac{\varphi(y)}{e} \right\} d\mu(y).$$

(So the operator T depends on φ.) However, their general theory requires that the operator T is weakly closed, which is an annoying technical condition. In our direct approach via the inequalities (1.10) and (1.11), no such extra conditions are necessary.

In this paper we work out the above in detail. In §2 we prove that the minimization problem (1.4) has a solution in $L^1(\Omega)$; in §3 we prove the inequalities (1.10) and (1.11). In §4 we derive the error and stability estimates, which imply that the method converges. Under some additional smoothness conditions on $\mathcal{K}$ we get asymptotic rates of convergence, and we prove a saturation result. In §5 we discuss the maximum entropy optimization problem (1.4) for finite-dimensional subspaces. In §6 we consider the maximum entropy method for approximate moment problems, and in §7 we discuss the (older) version of the maximum entropy problem for exact moment problems associated with integral equations in the style of Borwein and Lewis [2].

We finish the introduction with a remark on notation. In the remanider we let $\|\cdot\|$ denote the norm on the space $L^2(\Sigma)$ rather than the more cumbersome $\|\cdot\|_{L^2(\Sigma)}$, so for all $g \in L^2(\Sigma)$

$$\text{(1.18)}\qquad \|\, g\,\| = \left\{ \int_\Sigma |g(x)|^2\, d\mu(x) \right\}^{1/2}.$$

Norms on other $L^p(S)$ spaces are denoted explicitly as $\|\cdot\|_{L^p(S)}$.

2. Existence of regularized solutions.

For good measure we show that the maximum entropy problem (1.4)–(1.6) has a unique solution in $L^1(\Omega)$ for every $\alpha > 0$. The uniqueness follows from the strict convexity of $\|\mathcal{K}f - g\|^2 + \alpha^2 D(f, \varphi)$ as function of f. The essential ingredient in the existence proof is the weak compactness of sets of bounded entropy. Our proof is different from Borwein and Lewis [2] in that we avoid using the Kadec property of $D(\,\cdot\,, \varphi)$, but it is identical in spirit to that of Amato and Hughes [1] for the problem of minimizing the entropy functional $\int_\Omega f(y) \log f(y)\, dy$ subject to the constraint $\|\mathcal{K}f - g\| \leq \delta$. However, in the process we also prove that the solution is essentially positive, which is useful later on.

LEMMA 2.1. *The sets $\mathcal{E}_M$ defined as*

$$\mathcal{E}_M = \{\, f \in L^1(\Omega) \;:\; f \text{ nonnegative},\; D(f,\varphi) \le M \,\}, \tag{2.1}$$

are weakly closed, convex subsets of $L^1(\Omega)$.

Proof. The convexity of $\mathcal{E}_M$ follows from the convexity of $D(f,\varphi)$ as function of f. We show the closedness of $\mathcal{E}_M$ in the norm topology on $L^1(\Omega)$. Suppose that $\{f_n\}_n \subset \mathcal{E}_M$ converges strongly to $f_o \in L^1(\Omega)$. Then $f_n \longrightarrow f_o$ almost everywhere on Ω, and so $f_n \log(f_n/\varphi)+\varphi-f_n \longrightarrow f_o \log(f_o/\varphi)+\varphi-f_o$ almost everywhere on Ω. Since these are nonnegative functions, Fatou's lemma implies that $D(f_o,\varphi) \le \liminf_n D(f_n,\varphi) \le M$, so that $f_o \in \mathcal{E}_M$. Thus $\mathcal{E}_M$ is closed in $L^1(\Omega)$. The lemma now follows since closed convex sets are weakly closed; see Holmes [14, §12, Cor. 1]. □

COROLLARY 2.2. *For nonnegative $\varphi \in L^1(\Omega)$ the functional $f \longmapsto D(f,\varphi)$ is lower semicontinuous in the weak topology on $L^1(\Omega)$.*

Proof. Let $\{f_n\}_n \subset \mathcal{E}_M$ converge weakly to $f_o \in L^1(\Omega)$, and let $\{f_n'\}_n$ be a subsequence with $\lim_n D(f_n',\varphi) = \liminf_n D(f_n,\varphi)$. By Lemma 2.1 we have that $D(f_o,\varphi) \le \sup_n D(f_n',\varphi)$, and since this also holds for any subsequence of $\{f_n'\}_n$, it follows that $D(f_o,\varphi) \le \limsup_n D(f_n',\varphi) = \liminf_n D(f_n,\varphi)$, which shows the weak lower semicontinuity of $D(f,\varphi)$. □

Remark. Amato and Hughes [1] prove the weak lower semicontinuity of the functional $f \longmapsto \int_\Omega f \log f$ by appealing to Fatou's lemma, as in the proof of Lemma 2.1 above, but without explicit recourse to the convexity of this functional.

LEMMA 2.3. *For fixed nonnegative $\varphi \in L^1(\Omega)$, the sets $\mathcal{E}_M$ are weakly compact subsets of $L^1(\Omega)$.*

Proof. The compactness follows from the Dunford–Pettis criterion, combined with the criterion of De La Vallée–Poussin (see [8, Chap. 8, Thm. 3.1]) as follows. Assume without loss of generality that $\varphi(y) > 0$ almost everywhere on Ω. Define $m(y) = f(y)/\varphi(y)$. Then $f \in \mathcal{E}_M$ is equivalent with $m \in \mathcal{F}_M$, where $\mathcal{F}_M$ is defined as

$$\mathcal{F}_M = \left\{ m \in L^1(\Omega, \varphi(y)d\mu(y)) \;:\; \int_\Omega \Phi(m(y))\,\varphi(y)\,d\mu(y) \le M \right\}.$$

in which $\Phi(t) = t\log t + 1 - t$. Obviously, $\Phi(t)/t \longrightarrow \infty$ as $t \longrightarrow \infty$. Now the Dunford–Pettis–De La Vallée–Poussin criterion applies in the $L^1(\Omega,\varphi(y)dy)$-setting, and gives us the relative compactness of $\mathcal{F}_M$ in the weak topology on $L^1(\Omega,\varphi(y)d\mu(y))$. This is equivalent to the relative compactness of $\mathcal{E}_M$ in the weak topology on $L^1(\Omega)$. By Lemma 2.1, $\mathcal{E}_M$ is closed in the weak topology on $L^1(\Omega)$, so that the relative compactness of $\mathcal{E}_M$ implies its compactness. □

The above Corollary 2.2 and Lemma 2.3 immediately prove the existence of solutions to (1.4)–(1.6).

THEOREM 2.4. *For $\alpha > 0$, nonnegative $\varphi \in L^1(\Omega)$ and $g \in L^2(\Sigma)$ the maximum entropy problem* (1.4)–(1.6) *has a unique solution $f_\alpha \in L^1(\Omega)$. Moreover, the ratio f_α/φ is bounded away from zero.*

Proof. The functional $\ell(f) = \|\mathcal{K}f - g\|^2 + \alpha^2 D(f,\varphi)$ is weakly lower semicontinuous by Corollary 2.2 and the compactness of $\mathcal{K} : L^1(\Omega) \longrightarrow L^2(\Sigma)$. When minimizing $\ell(f)$ we may restrict attention to those f for which $\ell(f) \le \ell(\varphi)$, so that then $D(f,\varphi) \le M$ with $M = \alpha^{-1}\ell(\varphi)$. In other words, we may take $f \in \mathcal{E}_M$ as in (2.1). Thus we are considering a weakly lower semicontinuous functional on a weakly compact subset of $L^1(\Omega)$, and the existence of a solution follows. Note that f_α vanishes almost everywhere where φ vanishes.

Now suppose that $f_\alpha/\varphi \le \delta$ on some set $E \subset \Omega$ with positive measure (on which φ is positive almost everywhere), let $\mathbf{1}_E$ be the indicator function of the set E, and let $f_\lambda = f_\alpha + \lambda \mathbf{1}_E$. Consider the function $L(\lambda) = \|\mathcal{K} f_\lambda - g\|^2 + \alpha^2 D(f_\lambda, \varphi)$ for $\lambda \ge 0$. Then $L(\lambda)$ is differentiable, and

$$L'(0) = \int_E \left[\mathcal{K}^*(\mathcal{K} f_\alpha - g) + \alpha^2 \log \frac{f_\alpha}{\varphi}\right] \le \{\|\mathcal{K}^*(\mathcal{K} f_\alpha - g)\|_{C(\Omega)} + \alpha^2 \log \delta\}\, |E|.$$

So, if, $\delta < \exp\left(-\alpha^{-2}\|K^*(\mathcal{K} f_\alpha - g)\|_{C(\Omega)}\right)$, then $L'(0) \ge 0$, so that $L(\lambda)$ is not the minimum of $L(\lambda)$ for $\lambda \ge 0$. But, obviously, $L(\lambda)$ is minimal for $\lambda = 0$, so we have a contradiction, and f_α/φ is bounded away from 0. □

We note that it is not clear whether the solution f_α actually is in $L^2(\Omega)$ as well. The above analysis suggests a negative answer: the fact that $D(f, \varphi) \le M$ does not imply that $f \in L^2(\Omega)$.

3. Inequalities for constrained regularized least squares. In this section we prove the inequalities (1.10) and (1.11) for the regularized least squares problem (1.4), where $D(f, \varphi)$ has the representation (1.13). Since the nonnegativity constraint in (1.4) may appear rather special, we replace it with an abstract constraint $f \in \mathcal{C}$, where $\mathcal{C}$ is some closed convex set. Since the proper setting of the minimization problem (1.4) depends on the particular choice for D (e.g., for $D(f, \varphi) = \|f - \varphi\|^2_{L^2(\Omega)}$ the setting is the Hilbert space $L^2(\Omega)$, but for the entropy functional $D(f, \varphi)$ given by (1.6) the setting is $L^1(\Omega)$ as shown in §2), we let the setting be any Banach space X of integrable functions on Ω, and assume that $\mathcal{K} : X \longrightarrow L^2(\Sigma)$ is at least bounded. We now consider the problem (1.4) in slightly more abstract setting, i.e., we consider the problem

$$(3.1) \qquad \mathcal{P}_\alpha(g, \varphi): \quad \begin{array}{ll} \text{mimimize} & \ell(f) \overset{\text{def}}{=} \|\mathcal{K} f - g\|^2 + \alpha^2 D(f, \varphi), \\ \text{subject to} & f \in \mathcal{C}, \end{array}$$

where $\alpha > 0$, $g \in L^2(\Sigma)$, $\varphi \in \mathcal{C}$ and D given as

$$(3.2) \qquad D(f, \varphi) = d(f) - d(\varphi) - \langle d'(\varphi), f - \varphi \rangle.$$

Here $\langle \cdot, \cdot \rangle$ denotes the duality pairing between X and X^*, the dual of X, and d is a strictly convex, Gateaux differentiable function on $\mathcal{C}$, with $d'(f) \in X^*$ for each $f \in \mathcal{C}$. The following theorem is a result of Bregman [3] slightly modified. The corollary is vintage (varietal?) Bregman [3]. We remind the reader that $\|\cdot\|$ denotes the norm on $L^2(\Sigma)$.

THEOREM 3.1 (Bregman [3]). *Assuming that $\mathcal{P}_\alpha(g, \varphi)$ has a solution $f_\alpha \in \mathcal{C}$, we have for all $f \in \mathcal{C}$,*

$$\|\mathcal{K}(f - f_\alpha)\|^2 + \alpha^2 D(f, f_\alpha) \le \ell(f) - \ell(f_\alpha).$$

COROLLARY 3.2 (Bregman [3]). *Assuming that the problem*

$$(3.3) \qquad \textit{miminize} \quad D(f, \varphi) \quad \textit{subject to} \quad f \in \mathcal{C}$$

has a solution $f_\alpha \in \mathcal{C}$, we have for all $f \in \mathcal{C}$,

$$D(f, f_\alpha) \le D(f, \varphi) - D(f_\alpha, \varphi).$$

Proof of Corollary 3.2. In the theorem take $\mathcal{K} = \mathcal{O}$, the zero operator, and $g = 0$, the zero function. □

Proof of Theorem 3.1. We note that the representation (3.2) implies that

$$D(f,\varphi) - D(f_\alpha,\varphi) - D(f,f_\alpha) = \langle d'(f_\alpha) - d'(\varphi), f - f_\alpha\rangle = \langle D'(f,\varphi), f - f_\alpha\rangle.$$

A similar relation holds for $\|\mathcal{K}f - g\|^2$. Putting these two relations together results in

$$\ell(f) - \ell(f_\alpha) - \|\mathcal{K}(f - f_\alpha)\|^2 - \alpha^2 D(f,f_\alpha) = \langle \ell'(f_\alpha), f - f_\alpha\rangle,$$

which by the (necessary) Karash–Kuhn–Tucker conditions for a solution of $\mathcal{P}_\alpha(g,\varphi)$ is nonnegative for all $f \in \mathcal{C}$. □

It is possible to derive a continuity result for $\mathcal{P}_\alpha(g,\varphi)$, as follows.

THEOREM 3.3. *Let g, $G \in L^2(\Sigma)$. Let $f_\alpha \in \mathcal{C}$ solve the problem $\mathcal{P}_\alpha(g,\varphi)$, and let $F_\alpha \in \mathcal{C}$ solve the problem $\mathcal{P}_\alpha(G,\varphi)$. Then*

$$\|\mathcal{K}(f_\alpha - F_\alpha)\|^2 + \alpha^2 D(f_\alpha, F_\alpha) \le 4\,\|g - G\|^2.$$

Proof. From Theorem 3.1 we get that

$$\begin{aligned}(3.4)\quad &\|\mathcal{K}(f_\alpha - F_\alpha)\|^2 + \alpha^2 D(f_\alpha,F_\alpha)\\ &\qquad \le \|\mathcal{K}f_\alpha - G\|^2 + \alpha^2 D(f_\alpha,\varphi) - \|\mathcal{K}F_\alpha - G\|^2 - \alpha^2 D(F_\alpha,\varphi).\end{aligned}$$

A Taylor expansion in G around g gives

$$\|\mathcal{K}f_\alpha - G\|^2 - \|\mathcal{K}F_\alpha - G\|^2 = \|\mathcal{K}f_\alpha - g\|^2 - \|\mathcal{K}F_\alpha - g\|^2 + 2\langle \mathcal{K}(f_\alpha - F_\alpha), G - g\rangle,$$

so that from (3.4) we get that

$$\begin{aligned}(3.5)\quad &\|\mathcal{K}(f_\alpha - F_\alpha)\|^2 + \alpha^2 D(f_\alpha,F_\alpha) \le 2\,\|\mathcal{K}(f_\alpha - F_\alpha)\|\,\|g - G\|\\ &\qquad + \|\mathcal{K}f_\alpha - g\|^2 + \alpha^2 D(f_\alpha,\varphi) - \{\|\mathcal{K}F_\alpha - g\|^2 + \alpha^2 D(F_\alpha,\varphi)\}.\end{aligned}$$

Since f_α solves $\mathcal{P}_\alpha(g,\varphi)$, the expression on the second line of (3.5) is negative (nonpositive) so that

$$(3.6)\qquad \|\mathcal{K}(f_\alpha - F_\alpha)\|^2 + \alpha^2 D(f_\alpha,F_\alpha) \le 2\,\|\mathcal{K}(f_\alpha - F_\alpha)\|\,\|g - G\|.$$

Ignoring the $\alpha^2 D$ term on the left of (3.6) we get a quadratic inequality for $\|\mathcal{K}(f_\alpha - F_\alpha)\|$ with the result that

$$(3.7)\qquad \|\mathcal{K}(f_\alpha - F_\alpha)\| \le 2\,\|g - G\|.$$

Using (3.7) in the right-hand side of (3.6) then gives us the required result. □

For the general problem (3.3) an inequality as in Theorem 3.3 does not seem to be available, nor for the problem $\mathcal{P}_\alpha(g,\varphi)$ if we keep g fixed but allow φ to vary, let alone if we vary both g and φ. Note though that Theorem 3.3 in combination with inequality (1.7) already shows for the entropy case that

$$\|f_\alpha - F_\alpha\|_{L^1(\Omega)} \le c\,\|g - G\|,$$

for an appropriate c. For the entropy functional (1.6) we have the following result regarding continuous dependence on φ, but the type of variation allowed in φ is quite limited.

THEOREM 3.4. *Let D be given by (1.6). Let $g \in L^2(\Sigma)$, and let $\varphi \in L^1(\Omega)$ be nonnegative. Then there exists a constant c such that for all $G \in L^2(\Sigma)$ and $\Phi \in L^1(\Omega)$ for which $\|g - G\| + \|\log(\varphi/\Phi)\|_{L^\infty(\Omega)}$ is small enough*

$$\|\mathcal{K}(f_\alpha - F_\alpha)\|^2 + \alpha^2 D(f_\alpha, F_\alpha) \leq c \left(\| g - G \|^2 + \alpha^2 \| \log(\varphi/\Phi)\|^2_{L^\infty(\Omega)}\right),$$

where f_α solves $\mathcal{P}_\alpha(g, \varphi)$ and F_α solves $\mathcal{P}_\alpha(G, \Phi)$.

Proof. Since F_α solves $\mathcal{P}_\alpha(G, \Phi)$ it follows from Theorem 3.1 that

$$\begin{aligned}(3.8)\quad \|\mathcal{K}(f_\alpha - F_\alpha)\|^2 &+ \alpha^2 D(f_\alpha, F_\alpha) \\ &\leq \|\mathcal{K} f_\alpha - G\|^2 + \alpha^2 D(f_\alpha, \Phi) - \|\mathcal{K} F_\alpha - G\|^2 - \alpha^2 D(F_\alpha, \Phi),\end{aligned}$$

which leads to the following variation of (3.5):

$$\begin{aligned}\|\mathcal{K}(f_\alpha - F_\alpha)\|^2 + \alpha^2 D(f_\alpha, F_\alpha) &\leq 2\,\|\mathcal{K}(f_\alpha - F_\alpha)\|\,\|g - G\| \\ &+ \|\mathcal{K} f_\alpha - g\|^2 + \alpha^2 D(f_\alpha, \varphi) - \{\|\mathcal{K} F_\alpha - g\|^2 + \alpha^2 D(F_\alpha, \varphi)\} \\ &+ \alpha^2[D(f_\alpha, \Phi) - D(f_\alpha, \varphi) + D(F_\alpha, \varphi) - D(F_\alpha, \Phi)].\end{aligned} \tag{3.9}$$

Since f_α solves $\mathcal{P}_\alpha(g, \varphi)$ the second line of (3.9) is negative (nonpositive). The expression in the last line equals

$$\int_\Omega (f_\alpha - F_\alpha) \log \frac{\varphi}{\Phi} \leq \|f_\alpha - F_\alpha\|_{L^1(\Omega)}\, \|\log(\varphi/\Phi)\|_{L^\infty(\Omega)}.$$

By means of inequality (1.7) we may estimate

$$\|f_\alpha - F_\alpha\|_{L^1(\Omega)} \leq c\sqrt{D(f_\alpha, F_\alpha)},$$

for some constant c, so that

$$\int_\Omega (f_\alpha - F_\alpha) \log \frac{\varphi}{\Phi} \leq c\, \|\log(\varphi/\Phi)\|_{L^\infty(\Omega)}\, \sqrt{D(f_\alpha, F_\alpha)}.$$

Substituting this into (3.9) results in

$$E^2 + \alpha^2 D \leq 2\,E\,\|g - G\| + c\,\alpha^2 D^{1/2}\, \|\log(\varphi/\Phi)\|_{L^\infty(\Omega)},$$

where $E = \|\mathcal{K}(f_\alpha - F_\alpha)\|$, and $D = D(f_\alpha, F_\alpha)$. After the usual manipulations, as in the proof of Theorem 3.3, the estimate of the theorem follows. □

4. Convergence results and error estimates for maximum entropy. In this section we consider the convergence of the regularization method, and provide error estimates for maximum entropy. We will be using the results from §3 with $D(f, \varphi)$ the cross entropy given by (1.6). To motivate the following theorem it should be noted that for equations $\mathcal{K} f = g$ which admit nonnegative solutions its proof is rather straight forward; this suggests that (constrained) least squares solutions be admitted as well, and that is what we do. So we state first the *constrained least squares* (LS) problem:

$$\begin{aligned}(4.1)\qquad \text{LS:}\qquad &\text{minimize} \quad \|\mathcal{K} f - g\|^2 \\ &\text{subject to} \quad f \in L^1(\Omega), \quad f \geq 0,\end{aligned}$$

and then consider the *maximum entropy least squares* (MELS) problem

$$\text{(4.2)} \qquad \text{MELS:} \quad \begin{array}{ll} \text{minimize} & D(f,\varphi) \\ \text{subject to} & f \text{ solves LS.} \end{array}$$

The insistence on $L^1(\Omega)$ solutions to the LS problem is a nontrivial regularity condition. By way of example, we consider the case where k is bounded away from zero on $\Sigma \times \Omega$, i.e. $k(x,y) \geq k_{\min}$ for all x,y, for some positive constant $k_{\min}$. Then it follows for nonnegative $f \in L^1(\Omega)$ that

$$k_{\min} |\Sigma|^{1/2} \|f\|_{L^1(\Omega)} \leq \|\mathcal{K}f\| \leq \|g\| + \|\mathcal{K}f - g\|,$$

so that a minimizing sequence for (4.1) is bounded in $L^1(\Omega)$. Unfortunately, this does not rule out the possibility that the solution of (4.1) be a nonnegative measure on Ω which is not absolutely continuous with respect to Lebesgue measure. In order for (4.2) to have a solution we need to require even more than mere $L^1(\Omega)$-solvability of (4.1); cf. inequality (1.7).

THEOREM 4.1. *Assume that the* MELS *problem* (4.2) *has a solution* $f_o \in L^1(\Omega)$. *Let* $f_{\alpha\delta}$ *solve the problem* $\mathcal{P}_\alpha(g_\delta, \varphi)$, *with* $\|g_\delta - g\| \leq \delta$. *Then if* $\delta, \alpha \longrightarrow 0$, *with* $\delta/\alpha \longrightarrow 0$, *the regularized solutions* $f_{\alpha\delta}$ *converge strongly to* f_o, *i.e.,*

$$\|f_{\alpha\delta} - f_o\|_{L^1(\Omega)} \longrightarrow 0 \quad \textit{as } \delta \longrightarrow 0.$$

Proof. Applying Theorem 3.1 to $\mathcal{P}_\alpha(g_\delta, \varphi)$ we get that

$$\text{(4.3)} \quad \begin{aligned} &\|\mathcal{K}(f_o - f_{\alpha\delta})\|^2 + \alpha^2 D(f_o, f_{\alpha\delta}) \\ &\qquad \leq \|\mathcal{K}f_o - g_\delta\|^2 + \alpha^2 D(f_o, \varphi) - \|\mathcal{K}f_{\alpha\delta} - g_\delta\|^2 - \alpha^2 D(f_{\alpha\delta}, \varphi). \end{aligned}$$

By Taylor expansion (in g_δ around g) we get that

$$\|\mathcal{K}f_o - g_\delta\|^2 - \|\mathcal{K}f_{\alpha\delta} - g_\delta\|^2 = \|\mathcal{K}f_o - g\|^2 - \|\mathcal{K}f_{\alpha\delta} - g\|^2 + 2\langle \mathcal{K}(f_o - f_{\alpha\delta}), g_\delta - g\rangle.$$

The sum of the first two terms on the right is negative since f_o is a least squares solution. Straight forward estimation then gives that

$$\text{(4.4)} \qquad \|\mathcal{K}f_o - g_\delta\|^2 - \|\mathcal{K}f_{\alpha\delta} - g_\delta\|^2 \leq 2\delta E,$$

where $E = \|\mathcal{K}(f_{\alpha\delta} - f_o)\|$. Then we may write (4.3) as

$$\text{(4.5)} \qquad E^2 + \alpha^2 D(f_o, f_{\alpha\delta}) \leq 2\delta E + \alpha^2 \left[D(f_o, \varphi) - D(f_{\alpha\delta}, \varphi)\right].$$

Ignoring $\alpha^2 D$ term on the left-hand side of this inequality, as well as ignoring the term $-\alpha^2 D(f_{\alpha\delta}, \varphi)$ on the right, we get that $E^2 \leq 2\delta E + c\alpha^2$, where $c = D(f_o, \varphi)$. It follows that

$$\text{(4.6)} \qquad E \leq \delta + \sqrt{c\alpha^2 + \delta^2}.$$

For later use we note that from (4.6) it follows that E/α is bounded as α, $\delta \longrightarrow 0$, with $\delta/\alpha \longrightarrow 0$, and

$$\text{(4.7)} \qquad \|\mathcal{K}(f_{\alpha\delta} - f_o)\| \longrightarrow 0 \quad \text{for } \alpha, \delta \longrightarrow 0,$$

regardless of whether $\delta/\alpha \longrightarrow 0$ or not. Replacing the left-hand side of (4.5) by zero we get that

$$D(f_{\alpha\delta}, \varphi) \leq D(f_o, \varphi) + \frac{2\delta E}{\alpha^2}. \tag{4.8}$$

Since $\delta/\alpha \longrightarrow 0$ it follows from (4.6) that $\delta E/\alpha^2 \longrightarrow 0$, and so from (4.8) that

$$\limsup_{\alpha,\delta} D(f_{\alpha\delta}, \varphi) \leq D(f_o, \varphi). \tag{4.9}$$

Consequently, we have that the $f_{\alpha\delta}$ belong to a set of bounded entropy, so that by Lemma 2.3 the $f_{\alpha\delta}$ belong to a weakly compact subset of $L^1(\Omega)$. Now let $\alpha_n, \delta_n \longrightarrow 0$, with $\delta_n/\alpha_n \longrightarrow 0$ as $n \longrightarrow \infty$, let $g_n = g_{\delta_n}$, and $f_n = f_{\alpha_n\delta_n}$. Then $\{f_n\}_n$ has a weakly convergent subsequence, say with limit ψ_o. For notational convenience assume that the whole sequence converges weakly. Since $\mathcal{K} : L^1(\Omega) \longrightarrow L^2(\Sigma)$ is compact then $\mathcal{K}\psi_o = \lim_n \mathcal{K}f_n$ (strong limit in $L^2(\Sigma)$). From (4.7) we also have that $\mathcal{K}f_o = \lim_n \mathcal{K}f_n$ so that $\mathcal{K}\psi_o = \mathcal{K}f_o$. Since f_o solves the LS problem, then so does ψ_o. By the weak lower semicontinuity of $D(\,\cdot\,, \varphi)$ (see Corollary 2.2, and (4.9)) we also have that

$$D(\psi_o, \varphi) \leq \liminf_n D(f_n, \varphi) \leq \limsup_n D(f_n, \varphi) \leq D(f_o, \varphi).$$

Since f_o is the unique solution of MELS, and ψ_o solves LS, then $\psi_o = f_o$, and

$$\lim_n D(f_n, \varphi) = D(f_o, \varphi). \tag{4.10}$$

From (4.5), ignoring the E^2 term on the left, we obtain that

$$D(f_o, f_n) \leq \frac{2\delta_n E}{\alpha_n^2} + D(f_o, \varphi) - D(f_n, \varphi).$$

Since $\delta_n E/\alpha_n^2 \longrightarrow 0$ as discussed before, we get with (4.10) that $D(f_o, f_n) \longrightarrow 0$. The inequality (1.7) then shows that $\|f_n - f_o\|_{L^1(\Omega)} \longrightarrow 0$. This shows that every convergent subsequence of $\{f_{\alpha_n,\delta_n}\}_n$ converges to f_o. Consequently, the $f_{\alpha\delta}$ have only one accumulation point as $\alpha \longrightarrow 0$, $\delta/\alpha \longrightarrow 0$, and the theorem follows. □

The above shows that the regularized solutions converge, but a *rate of convergence* is not provided. Inspection of the above proof reveals that in order to get rates we need to estimate the term $\alpha^2[D(f_o, \varphi) - D(f_{\alpha\delta}, \varphi)]$ in (4.5). From the easily checked identity

$$D(f_o, \varphi) - D(f_{\alpha\delta}, \varphi) = -D(f_{\alpha\delta}, f_o) + \int_\Omega (f_o - f_{\alpha\delta}) \log(f_o/\varphi),$$

which holds provided f_o/φ is nonzero almost everywhere, we see that the assumption

$$\log(f_o/\varphi) = \mathcal{K}^* \varrho_o, \tag{4.11}$$

for some $\varrho_o \in L^2(\Sigma)$ would go a long way towards this goal. At the end of this section we will address the question whether this assumption is at all reasonable.

THEOREM 4.2. *Assume that f_o solves the* MELS *problem and satisfies* (4.11). *Let $f_{\alpha\delta}$ solve the problem $\mathcal{P}_\alpha(g_\delta,\varphi)$, with $\|g_\delta - g\| \leq \delta$. Then for $\delta \longrightarrow 0$ with $\alpha = \delta^{1/2}$, the regularized solutions $f_{\alpha\delta}$ converge to f_o, and*

$$\|f_{\alpha\delta} - f_o\|_{L^1(\Omega)} = \mathcal{O}(\delta^{1/2}), \quad \delta \longrightarrow 0.$$

Proof. Our starting point is (4.5) in the proof of Theorem 4.1. For the relevant terms in the right-hand side of (4.5) we get by convexity that

$$D(f_o,\varphi) - D(f_{\alpha\delta},\varphi) \leq \int_\Omega (f_o - f_{\alpha\delta}) \log \frac{f_o}{\varphi},$$

so that by (4.11)

$$D(f_o,\varphi) - D(f_{\alpha\delta},\varphi) \leq \int_\Omega (f_o - f_{\alpha\delta})\mathcal{K}^* \varrho_o = \int_\Sigma \varrho_o \mathcal{K}(f_o - f_{\alpha\delta}).$$

It follows that for some constant c (depending on f_o!),

$$\alpha^2 \left[D(f_o,\varphi) - D(f_{\alpha\delta},\varphi)\right] \leq c\,\alpha^2 E. \tag{4.12}$$

Substituting the estimate (4.12) into the right-hand side of (4.5) gives for some (other) constant c,

$$E^2 + \alpha^2 D(f_o, f_{\alpha\delta}) \leq c\,(\delta + \alpha^2)E. \tag{4.13}$$

Ignoring the $\alpha^2 D$ term on the left yields $E \leq c\,(\delta + \alpha^2)$, so that now ignoring the E^2 on the left in (4.13) gives $D(f_o, f_{\alpha\delta}) \leq [c\,(\alpha + \delta/\alpha)]^2$. The choice $\alpha = \delta^{1/2}$ minimizes the right-hand side, and finally (1.7) gives the required result. □

We have an analogous result for exact data (i.e., $\delta = 0$)

THEOREM 4.3. *Assume that f_o solves the* MELS *problem* (4.1) *and satisfies* (4.11). *Let f_α solve $\mathcal{P}_\alpha(g,\varphi)$. Then as $\alpha \longrightarrow 0$ the regularized solution f_α converges to f_o and*

$$D(f_\alpha, f_o) = \mathcal{O}(\alpha^2), \qquad \|\mathcal{K}(f_\alpha - f_o)\| = \mathcal{O}(\alpha^2),$$

as well as

$$\|f_\alpha - f_o\|_{L^1(\Omega)} = \mathcal{O}(\alpha).$$

Proof. As in the proof of the previous theorem we get the inequality (4.13) for $E = \|\mathcal{K}(f_\alpha - f_o)\|$, but now with $\delta = 0$, so that

$$E^2 + \alpha^2 D(f_o, f_\alpha) \leq c\,\alpha^2 E,$$

from which it follows that $E \leq c\,\alpha^2$, and then $D(f_o, f_\alpha) \leq c^2\alpha^2$, and (1.7) does the trick. □

We now show the converse of Theorem 4.3. Although this does not show the reasonableness per se of condition (4.11), it does show that the above error estimates are possible *only* if (4.11) holds. This type of theorem is known as a *saturation* result; cf. Groetsch [12].

THEOREM 4.4. *Let $\varphi \in L^1(\Omega)$ be nonnegative, and let $g \in L^2(\Sigma)$. Let f_o be the solution of the* MELS *problem, and assume that $\mathcal{K}^*(\mathcal{K}f_o - g) = 0$. Let f_α be the solution of $\mathcal{P}_\alpha(g, \varphi)$, and suppose that*

$$D(f_\alpha, f_o) = \mathcal{O}(\alpha^2), \qquad \|\mathcal{K}(f_\alpha - f_o)\| = \mathcal{O}(\alpha^2).$$

Then condition (4.11) *holds.*

Proof. Since f_α solves $\mathcal{P}_\alpha(g, \varphi)$, and since f_α is bounded away from zero (see Theorem 2.4) it follows that the gradient of the objective function vanishes, so

$$2\mathcal{K}^*\mathcal{K}(f_\alpha - f_o) + \alpha^2 \log \frac{f_\alpha}{\varphi} = 0 \quad \text{a.e. on } \Omega, \tag{4.14}$$

where we used the assumption that $\mathcal{K}^*g = \mathcal{K}^*\mathcal{K}f_o$. Let $r_\alpha = -2\mathcal{K}(f_\alpha - f_o)/\alpha^2$, then $\|r_\alpha\|$ is bounded as $\alpha \longrightarrow 0$, by our assumptions. Then for a sequence $\{\alpha_n\}_n$, with $\alpha_n \longrightarrow 0$, we have that $\{r_{\alpha_n}\}_n$ converges weakly to some element $r_o \in L^2(\Sigma)$. Then $\{\mathcal{K}^*r_{\alpha_n}\}_n$ converges strongly to $\mathcal{K}^*r_o$ in the $C(\Omega)$-topology. It follows from (4.14) that then $\log(f_o/\varphi) = \mathcal{K}^*r_o$, and the theorem follows. □

5. Finite-dimensional approximation of maximum entropy problems. In this section we consider the convergence of the solutions of the maximum entropy problems when the maximization is done over finite-dimensional subspaces. So we let $V_1 \subset V_2 \subset \cdots \subset V_n \subset \cdots$ be a nested sequence of subspaces of $L^1(\Omega)$, such that $\bigcup_n V_n$ is dense in $L^1(\Omega)$. We consider the convergence of the solution f_n of the problem

$$\begin{array}{ll} \text{mimimize} & \ell(f) \overset{\text{def}}{=} \|\mathcal{K}f - g\|^2 + \alpha^2 D(f, \varphi), \\ \text{subject to} & f \in V_n, \ f \geq 0, \end{array} \tag{5.1}$$

for fixed $\alpha > 0$ to the solution f_α of the regularized least squares problem (1.4)–(1.6), as well as the convergence for $\alpha = \alpha_n . \longrightarrow 0$ to the solution f_o to the MELS problem (4.2). We need a suitable, possibly nonlinear projector mapping nonnegative $L^1(\Omega)$ functions into nonnegative elements of V_n. We denote this projector by q_n. We require q_n to satisfy the following approximation property:

$$\begin{array}{l} q_n \text{ maps nonnegative } L^1(\Omega) \text{ functions into nonnegative functions in } V_n \\ \text{such that } \lim\limits_{n \longrightarrow \infty} D(q_n f, f) = 0 \text{ for all nonnegative } f \in L^1(\Omega). \end{array} \tag{5.2}$$

Later we will discuss the existence of such q_n.

THEOREM 5.1. *Let $\alpha > 0$ be fixed, let $\varphi \in L^1(\Omega)$ be nonnegative, and let $g \in L^2(\Sigma)$. Let f_n denote the solution of* (5.1). *Then $\{f_n\}_n$ converges to the solution f_α of* (3.1) *and*

$$\|\mathcal{K}(f_n - f_\alpha)\|^2 + \alpha^2 D(f_n, f_\alpha) \leq \|\mathcal{K}(q_n f_\alpha - f_\alpha)\|^2 + \alpha^2 D(q_n f_\alpha, f_\alpha).$$

Proof. Theorem 3.1 is the starting point as always. Since f_α solves $\mathcal{P}_\alpha(g, \varphi)$, we get that

$$\|\mathcal{K}(f_n - f_\alpha)\|^2 + \alpha^2 D(f_n, f_\alpha) \leq \ell(f_n) - \ell(f_\alpha). \tag{5.3}$$

Using the identities

$$-\|\mathcal{K}f_\alpha - g\|^2 = -\|\mathcal{K}q_n f_\alpha - g\|^2 + 2\langle \mathcal{K}f_\alpha - g, \mathcal{K}(q_n f_\alpha - f_\alpha)\rangle + \|\mathcal{K}(q_n f_\alpha - f_\alpha)\|^2,$$

and

$$-D(f_\alpha, \varphi) = -D(q_n f_\alpha, \varphi) + D(q_n f_\alpha, f_\alpha) - \langle D'(f_\alpha, \varphi), f_\alpha - q_n f_\alpha\rangle,$$

we get that

$$\begin{aligned} &\|\mathcal{K}(f_n - f_\alpha)\|^2 + \alpha^2 D(f_n, f_\alpha) \leq \ell(f_n) - \ell(q_n f_\alpha) \\ &\qquad + \|\mathcal{K}(f_\alpha - q_n f_\alpha)\|^2 + \alpha^2 D(q_n f_\alpha, f_\alpha) \\ &\qquad + 2\langle \mathcal{K}^*(\mathcal{K} f_\alpha - g), q_n f_\alpha - f_\alpha\rangle + \alpha^2\langle D'(f_\alpha, \varphi), q_n f_\alpha - f_\alpha\rangle. \end{aligned} \tag{5.4}$$

The expression on the right in the first line of (5.4) is negative (nonpositive) since f_n solves (3.1), and $q_n f_\alpha \in V_n$ as well. The expression on the third line of (5.4) equals $\langle \ell'(f_\alpha), q_n f_\alpha - f_\alpha\rangle$. Now note that the fact that $D(q_n f_\alpha, f_\alpha)$ is finite implies that $q_n f_\alpha$ vanishes on the set where f_α itself vanishes. Since $\ell'(f_\alpha)$ and f_α cannot both be positive on the same sets of positive measure, it follows that the same holds for $\ell'(f_\alpha)$ and $q_n f_\alpha$. Consequently

$$\langle \ell'(f_\alpha), q_n f_\alpha - f_\alpha\rangle = 0,$$

so that from (5.4) the result follows. □

The inequality of the theorem is quite satisfactory, since the error is related to how well f_α can be approximated by elements from V_n. We are also interested in the case $\alpha = \alpha_n \longrightarrow 0$. To quantify the choice of α_n we need the quantity ε_n defined as

$$\varepsilon_n^2 = \|\mathcal{K}(q_n f_\alpha - f_\alpha)\|^2 + \alpha^2 D(q_n f_\alpha, f_\alpha). \tag{5.5}$$

By the assumption (5.2) and inequality (1.7) we have that $\varepsilon_n \longrightarrow 0$ as $n \longrightarrow \infty$.

THEOREM 5.2. *Let $\varphi \in L^1(\Omega)$ be nonnegative, let $g \in L^2(\Sigma)$, and let $\alpha = \alpha_n$ with $\alpha_n = \varepsilon_n^{1/2}$. Assume that the solution f_o of the* MELS *problem satisfies the condition* (4.11). *Then $\{f_n\}_n$ converges to the solution f_o, and*

$$\|f_n - f_o\|_{L^1(\Omega)} \leq c\,\varepsilon_n^{1/2},$$

for some constant c.

Proof. Let ψ_n denote the solution of (3.1) for $\alpha = \alpha_n$. So $\psi_n = f_{\alpha_n}$. By Theorem 5.1 and inequality (1.7) we have that $\|f_n - \psi_n\|_{L^1(\Omega)} \leq c\,\varepsilon_n/\alpha_n$, and by Theorem 4.3 we have that $\|\psi_n - f_o\|_{L^1(\Omega)} \leq c\,\alpha_n$, so that by the triangle inequality

$$\|f_n - f_o\|_{L^1(\Omega)} \leq c\,(\alpha_n + \varepsilon_n/\alpha_n),$$

for some (other) constant c. Since $\alpha_n = \varepsilon_n^{1/2}$ the result follows. □

We finish this section with a discussion of the nonlinear projection operator q_n and the associated approximation number ε_n, see (5.5). The best choice (to produce the smallest ε_n) would be to define $q_n f$ as the solution ψ of the minimization problem

$$\begin{aligned} &\text{mimimize} \quad \|\mathcal{K}(\psi - f)\|^2 + \alpha^2 D(\psi, f) \\ &\text{subject to} \quad \psi \in V_n\,,\ \psi \geq 0. \end{aligned} \tag{5.6}$$

Theorem 5.1 implies that the optimal ψ for $f = f_\alpha$ is precisely equal to f_n, the solution of (5.1), and so implies the commutativity of the diagram

$$\begin{tikzcd}
(g,\varphi) \arrow[r, "(3.1)"] \arrow[dr, "(5.1)"'] & f_\alpha \arrow[d, "(5.6)"] \\
& f_n
\end{tikzcd}$$

So to show that $\varepsilon_n \longrightarrow 0$ it suffices to indicate a $\psi_n \in V_n$ such that $\eta_n \longrightarrow 0$, where

$$\eta_n = \|\mathcal{K}(\psi_n - f)\|^2 + \alpha^2 D(\psi_n, f). \tag{5.7}$$

By inequality (1.7) all we need to do is show that $D(\psi_n, f) \longrightarrow 0$ as $n \longrightarrow \infty$. We will show this for one special choice of the subspaces V_n. Let $\{\sigma_{in}\}_i$ be a nested sequence of triangulations of Ω, with maximum diameter tending to zero as $n \longrightarrow \infty$, i.e., with $\operatorname{diam}(\sigma) = \sup\{|x-y| \,:\, x, y \in \sigma\}$ we require that

$$\max_i \operatorname{diam}(\sigma_{in}) \longrightarrow 0 \quad (n \longrightarrow \infty). \tag{5.8}$$

Let p be a positive natural number. We let V_n be the space of $C^{(-1)}$ piecewise polynomial functions of degree p or less. So $\psi \in V_n$ if and only if ψ restricted to σ_{in} is a polynomial of degree p or less. No continuity conditions across boundaries of the σ_{in} are imposed.

We proceed to construct the ψ_n for given nonnegative $f \in L^1(\Omega)$. Let $p_n : L^1(\Omega) \longrightarrow V_n$ be a projection operator, e.g., the choice

$$p_n f(y) = \tfrac{1}{|\sigma_{in}|} \int_{\sigma_{in}} f(z)\, d\mu(z) \quad \text{for } y \in \sigma_{in}, \tag{5.9}$$

would do. Set $\zeta_n = \|f - p_n f\|_{L^1(\Omega)}$, then $\zeta_n \longrightarrow 0$ as $n \longrightarrow 0$. Define the sets A_n, B_n by

$$A_n = \{\, y \in \Omega \,:\, f(y) < \zeta_n \,\}, \quad B_n = \{\, y \in \Omega \,:\, |f(y) - p_n f(y)| > \zeta_n^{2/3} \,\}. \tag{5.10}$$

Then $|B_n|$, the measure of B_n, tends to zero as $n \longrightarrow \infty$. Now define $\psi_n(y)$ for $y \in \sigma_{in}$ as

$$\psi_n(y) = \begin{cases} 0, & \text{if } \sigma_{in} \cap (A_n \cup B_n) \neq \emptyset, \\ p_n f(y) - m_{in}, & \text{otherwise,} \end{cases} \tag{5.11}$$

where $m_{in} = 0$ if $p_n f(y)$ is nonnegative on σ_{in}, and $m_{in} = \min\{p_n f(y) \,:\, y \in \sigma_{in}\}$ otherwise. So ψ_n is nonnegative everywhere. Also note that $|m_{in}| \le \zeta_n^{2/3}$. We now estimate the quantity

$$\int_\Omega \frac{|\psi_n(y) - f(y)|^2}{f(y)}\, d\mu(y) = \left[\int_{A_n} + \int_{B_n} + \int_{C_n}\right] \frac{|\psi_n(y) - f(y)|^2}{f(y)}\, d\mu(y), \tag{5.12}$$

where $C_n = \Omega \setminus (A_n \cup B_n)$. First we have that

$$\int_{A_n} \frac{|\psi_n(y) - f(y)|^2}{f(y)}\, d\mu(y) = \int_{A_n} f(y)\, d\mu(y) \le \zeta_n\, |A_n| \le \zeta_n\, |\Omega| \longrightarrow 0 \quad (n \longrightarrow \infty),$$

as well as

$$\int_{B_n} \frac{|\psi_n(y) - f(y)|^2}{f(y)}\, d\mu(y) = \int_{B_n} f(y)\, d\mu(y) \longrightarrow 0,$$

since $|B_n| \longrightarrow 0$ as $n \longrightarrow \infty$.

Secondly, on C_n we have that $f(y) \geq \zeta_n$ and $|\psi_n(y) - f(y)| \leq |p_n f(y) - f(y)| + |m_{in}|$, so that

$$\int_{C_n} \frac{|\psi_n(y) - f(y)|^2}{f(y)}\, d\mu(y) \leq \int_{C_n} \frac{(\zeta_n^{2/3} + |m_{in}|)^2}{\zeta_n} \leq 4\,\zeta_n^{1/3}\,|\Omega| \longrightarrow 0 \quad (n \longrightarrow \infty).$$

Putting the above together shows that

$$\int_{\Omega} \frac{|\psi_n(y) - f(y)|^2}{f(y)}\, d\mu(y) \longrightarrow 0 \quad (n \longrightarrow \infty),$$

and then from (1.8) we have that $D(\psi_n, f) \longrightarrow 0$ as desired.

The above argument becomes a bit more involved if continuity/differentiability conditions are imposed on the elements of V_n across the boundaries of the triangulations. It seems reasonable to assume that the conclusion would remain the same.

6. Convergence results for moment problems I. In this section we consider maximum entropy regularization for approximate moment problems, i.e., we assume that we have available for some nonnegative $f_o \in L^1(\Omega)$ the data g_{in} with

$$g_{in} = [\mathcal{K} f_o](x_i) + d_{in}, \qquad i = 1, 2, \ldots, n, \tag{6.1}$$

where d_{in} represents noise. The *exact* moment problem in the style of Borwein and Lewis [2] is considered in §7. Here we regularize (6.1) by means of the minimization problem

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{i=1}^{n} w_{in} \left|[\mathcal{K} f](x_i) - g_{in}\right|^2 + \alpha^2 D(f, \varphi) \\ \text{subject to} & f \in L^1(\Omega), \quad f \geq 0, \end{array} \tag{6.2}$$

for appropriate weights w_{in}, about which more later on. In this setting we run into trouble with the point evaluation functionals $\mathcal{K} f \longmapsto [\mathcal{K} f](x_i)$, which are not bounded in the $L^2(\Sigma)$ setting, viz. there is no constant c such that

$$\left|[\mathcal{K} f](x_i)\right| \leq c\,\|\mathcal{K} f\| \quad \text{for all } f \in L^1(\Omega). \tag{6.3}$$

When $D(f, \varphi) = \|f - \varphi\|_{\mathcal{H}}^2$, where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) there are quite elegant solutions to this; see Natterer [20] and Lukas [18]. A similar case arises when we assume that g and $\mathcal{K} f$ themselves lie in some RKHS; see Nashed and Wahba [19], and (much) more recently Nychka and Cox [21]. Here we fudge the issue by considering approximations to the point evaluations in the form

$$r_n F(i) = \int_{\Sigma} F(x) r_{in}(x)\, d\mu(x), \quad i = 1, 2, \ldots, n \tag{6.4}$$

for all $F \in C(\Sigma)$, and we require an (approximate) interpolation operator p_n such that

$$\|F - p_n r_n F\|_{C(\Sigma)} \longrightarrow 0 \quad \text{as } n \longrightarrow \infty \tag{6.5}$$

for all $F \in C(\Sigma)$. We set $P_n = p_n r_n$ and we define the weights w_{in} by means of

$$\sum_{i=1}^{n} w_{in}\,[r_n F](i) = \int_\Sigma [P_n F](x)\,d\mu(x) \quad \text{for all } F \in C(\Sigma). \tag{6.6}$$

We now rephrase (6.2) as

$$\begin{array}{ll} \text{minimize} & \ell_n(f) \overset{\text{def}}{=} \|P_n(\mathcal{K}f - g_n)\|^2 + \alpha^2 D(f,\varphi) \\ \text{subject to} & f \in L^1(\Omega),\ f \geq 0, \end{array} \tag{6.7}$$

with $g_n = p_n(\{g_{in}\}_{i=1}^n)$.

In the analysis below it is helpful to have the following quantities available,

$$\begin{aligned} \varepsilon_n &= \sup\{\|(I - P_n^*)\mathcal{K}f\| + \|(I - P_n)\mathcal{K}f\| \;:\; \|f\|_{L^1(\Omega)} \leq 1\}, \\ \eta_n &= \|(I - P_n^*)g\| + \|(I - P_n)g\|, \\ \delta_n &= \|P_n(g - g_n)\|. \end{aligned} \tag{6.8}$$

We assume that $\varepsilon_n, \eta_n \longrightarrow 0$ as $n \longrightarrow \infty$. Note the occurence of the operator $I - P_n^*$, so that this requires more than just (6.5). We also assume that $\delta_n \longrightarrow 0$.

We begin the analysis with a boundedness result.

LEMMA 6.1. *Let f_n be the solution of* (6.7), *and f_α the solution of* (3.1). *Then*

$$\begin{aligned} D(f_n,\varphi) &\leq D(f_o,\varphi) + \delta_n^2/\alpha^2, \\ D(f_\alpha,\varphi) &\leq D(f_o,\varphi). \end{aligned}$$

Proof. Since f_n solves (6.7) we have that $\ell_n(f_n) \leq \ell_n(f_o)$, so that ignoring the term $\|P_n(\mathcal{K}f_n - g_n)\|$ on the left we get

$$\alpha^2 D(f_n,\varphi) \leq \alpha^2 D(f_o,\varphi) + \|P_n(\mathcal{K}f_o - g_n)\|^2.$$

Since $\mathcal{K}f_o = g$, this proves the inequality. The argument for f_α proceeds in the same way. □

COROLLARY 6.2. *There exists a constant c such that*

$$\|f_n\|_{L^1(\Omega)} \leq c\left(1 + \frac{\delta_n}{\alpha}\right), \qquad \|f_\alpha\|_{L^1(\Omega)} \leq c.$$

Proof. The proof follows from the previous lemma, inequality (1.7), and Lemma 2.3. □

We are now ready to consider the convergence of the entropy method.

THEOREM 6.3. *Let $\varphi \in L^1(\Omega)$ be nonnegative, let g, $g_n \in L^2(\Sigma)$. Let f_n denote the solution of* (6.7). *Then $\{f_n\}_n$ converges to the solution f_α of* (3.1) *and*

$$\|f_n - f_\alpha\|_{L^1(\Omega)} \leq c\,\alpha^{-1}\left(1 + \frac{\delta_n}{\alpha}\right)(\eta_n + \delta_n + \varepsilon_n).$$

Proof. Since f_n solves (6.7) we have that $0 \leq \ell_n(f_\alpha) - \ell_n(f_n)$, and since f_α solves (3.1) we have again by Theorem 3.1 that

$$\|\mathcal{K}(f_n - f_\alpha)\|^2 + \alpha^2 D(f_n, f_\alpha) \leq \ell(f_n) - \ell(f_\alpha).$$

Adding both inequalities results in

$$\begin{aligned}&\|\mathcal{K}(f_n - f_\alpha)\|^2 + \alpha^2 D(f_n, f_\alpha)\\ &\quad\le \|P_n(\mathcal{K}f_\alpha - g_n)\|^2 - \|\mathcal{K}f_\alpha - g\|^2 + \|\mathcal{K}f_n - g\|^2 - \|P_n(\mathcal{K}f_n - g_n)\|^2.\end{aligned}$$

In the following we let $F_\alpha = \mathcal{K}f_\alpha - g$, $F_n = \mathcal{K}f_n - g$, and $\gamma_n = P_n(g_n - g)$, as well as $Q_n = P_n - I$. Then the above may be rewritten as

$$\begin{aligned}&\|\mathcal{K}(f_n - f_\alpha)\|^2 + \alpha^2 D(f_n, f_\alpha)\\ &\quad\le \|P_nF_\alpha - \gamma_n\|^2 - \|F_\alpha\|^2 + \|F_n\|^2 - \|P_nF_n - \gamma_n\|^2.\end{aligned}$$

By Taylor expansion of the right-hand side this results in

$$\begin{aligned}&\le 2\langle F_\alpha, Q_nF_\alpha - \gamma_n\rangle + \|Q_nF_\alpha - \gamma_n\|^2 - 2\langle F_n, Q_nF_n - \gamma_n\rangle - \|Q_nF_n - \gamma_n\|^2\\ &\le 2\langle F_\alpha, Q_n\mathcal{K}(f_\alpha - f_n)\rangle - 2\langle \mathcal{K}(f_n - f_\alpha), Q_nF_n - \gamma_n\rangle\\ &\quad + \|Q_nF_\alpha - \gamma_n\|^2 - \|Q_nF_n - \gamma_n\|^2.\end{aligned}$$

Writing $\langle F_\alpha, Q_n\mathcal{K}(f_\alpha - f_n)\rangle = \langle Q_n^*F_\alpha, K(f_\alpha - f_n)\rangle$, this results in the inequality

$$E^2 + \alpha^2 D(f_n, f_\alpha) \le 2\,E\,A_n + B_n^2, \tag{6.9}$$

where $E = \|\mathcal{K}(f_n - f_\alpha)\|$, and

$$\begin{aligned}B_n &= \|Q_nF_\alpha - \gamma_n\|,\\ A_n &= \|Q_n^*F_\alpha\| + \|Q_nF_n\| + \|\gamma_n\|.\end{aligned} \tag{6.10}$$

We may rewrite (6.9) as $(E - A_n)^2 + \alpha^2 D(f_n, f_\alpha) \le A_n^2 + B_n^2$, so that, ignoring the first term on the left we get that

$$\alpha^2 D(f_n, f_\alpha) \le A_n^2 + B_n^2.$$

Since we may estimate A_n and B_n as

$$\begin{aligned}B_n &\le \varepsilon_n\|f_n\|_{L^1(\Omega)} + \eta_n + \delta_n,\\ A_n &\le (\|f_\alpha\|_{L^1(\Omega)} + \|f_n\|_{L^1(\Omega)})\,\varepsilon_n + \eta_n + \delta_n\ .\end{aligned}$$

Corollary 6.2 and inequality (1.7) then provide us with the required estimate. □

We now have a result analogous to Theorem 4.2.

THEOREM 6.4. *Assume that* $\mathcal{K}f_o = g$ *and that* f_o *satisfies condition* (4.11). *Let* $e_n = \varepsilon_n + \eta_n + \delta_n$. *Let* f_n *solve* (6.2) *with* $\alpha = \alpha_n$, *where* $\alpha_n = e_n^{1/2}$. *Then*

$$\|f_n - f_\alpha\|_{L^1(\Omega)} \le c\,e_n^{1/2}.$$

Proof. We have that $\|f_n - f_o\| \le \|f_n - f_\alpha\| + \|f_\alpha - f_o\|$. By Theorem 6.3 and Theorem 4.3 we then have that

$$\|f_n - f_o\|_{L^1(\Omega)} \le c\left(\alpha_n + \frac{e_n}{\alpha_n}\left(1 + \frac{\delta_n}{\alpha_n}\right)\right) \le \tilde{c}\left(\alpha_n + \frac{e_n}{\alpha_n}\right),$$

the last inequality owing to the fact that $\delta_n/\alpha_n \longrightarrow 0$. The theorem follows. □

Specific error estimates follow once the operators r_n, p_n and P_n have been chosen, along the lines of Lukas [18], Nychka and Cox [21], or Nashed and Wahba [19]. It should be noted that here these operators play a crucial part in the method, as opposed to the operator q_n in §5, which played a *theoretical* part only. At this point we also mention that we *can* avoid the introduction of the approximate point evaluation functionals r_n (see (6.4)), but at the price of an extra factor α^{-1} in the estimate of Theorem 6.3, similar to the proof of Theorem 7.2 below. We omit the details.

7. Convergence results for moment problems II. In this section we consider the convergence of the maximum entropy method for the moment problem for $\mathcal{K}f = g$, inspired by Borwein and Lewis [2],

$$\begin{aligned} &\text{mimimize} \quad D(f,\varphi) \\ &\text{subject to} \quad f \geq 0, \quad [\mathcal{K}f](x_i) = g_i, \quad i = 1,2,\dots,n, \end{aligned} \tag{7.1}$$

where the $x_i \in \Sigma$ are given, and the data consists of the $g_i = g(x_i)$ and $\varphi \in L^1(\Omega)$. We are interested in the behavior of the solution f_n as $n \longrightarrow \infty$. Borwein and Lewis [2] obtain uniform convergence for some classical moment problems under quite natural conditions. Here we will be satisfied with obtaining $L^1(\Omega)$ convergence. It should be noted that as a regularization method the method (7.1) is rather suspect, the biggest problem being that the constraints with noisy data

$$[\mathcal{K}f](x_i) = g(x_i) + \delta_i, \qquad i = 1,2,\dots,n, \tag{7.2}$$

might easily be *inconsistent.* For consistent problems we have the following result.

THEOREM 7.1. *Let $\varphi \in L^1(\Omega)$ be nonnegative, and let $\{g_i\}_i$ be a bounded sequence of real numbers. Suppose that there exists a nonnegative $f_o \in L^1(\Omega)$ such that $D(f_o,\varphi) < \infty$, and*

$$[\mathcal{K}f_o](x_i) = g_i, \quad i = 1,2,\dots. \tag{7.3}$$

Let $f_n \in L^1(\Omega)$ be the solution of (7.1)*. Then $\{f_n\}_n$ converges strongly to the solution of*

$$\begin{aligned} &\textit{mimimize} \quad D(f,\varphi) \\ &\textit{subject to} \quad f \geq 0, \ [\mathcal{K}f](x_i) = g_i, \quad i = 1,2,\dots. \end{aligned} \tag{7.4}$$

Proof. Let $\mathcal{C}_n \subset L^1(\Omega)$ be the set of nonnegative $f \in L^1(\Omega)$, which satisfy

$$[\mathcal{K}f](x_i) = g_i, \quad i = 1,2,\dots,n.$$

Note that $f_o \in \mathcal{C}_n$ so that $\mathcal{C}_n$ is nonempty, and obviously the $\mathcal{C}_n$ are closed and convex subsets of $L^1(\Omega)$, and $\mathcal{C}_m \subset \mathcal{C}_n$ if $m > n$. Then the solution f_n to (7.1) exists by Theorem 2.4, and for $m > n$ we have $f_m \in \mathcal{C}_n$. By Corollary 3.2 we have that

$$D(f_m,f_n) \leq D(f_m,\varphi) - D(f_n,\varphi) \quad \text{for all } m > n. \tag{7.5}$$

Since $f_o \in \mathcal{C}_n$ then $D(f_n,\varphi) \leq D(f_o,\varphi)$, so it follows that $\{D(f_n,\varphi)\}_n$ is a bounded increasing sequence, hence it has a finite limit. Then the right-hand side of (7.5) tends to zero as $n, m \longrightarrow \infty$, and we then have that $D(f_m,f_n) \longrightarrow 0$ as n, $m \longrightarrow \infty$ (with $m \geq n$). From the inequality (1.7) we then get that $\{f_n\}_n$ is Cauchy in $L^1(\Omega)$, and thus is convergent, say with limit ψ_o. It is obvious that ψ_o satisfies (7.3). Since $f_o \in \mathcal{C}_n$ it follows from Corollary 3.2 that

$$0 \leq D(f_o,f_n) \leq D(f_o,\varphi) - D(f_n,\varphi),$$

and so $D(\psi_o,\varphi) = \lim_n D(f_n,\varphi) \leq D(f_o,\varphi)$. Since this holds for all f_o which satisfy (7.3), this shows that ψ_o solves (7.4). □

It is clear that we are a long way from uniform convergence, and we will not try to recreate such results. See Borwein and Lewis [2]. However, under the usual conditions (4.11) it is possible to get L^1-error estimates. So we assume that $\{x_i\}_i$ is dense in Σ, with

$$\sup_{y\in\Sigma}\min_{1\leq i\leq n}|y-x_i|\longrightarrow 0,\quad n\longrightarrow\infty. \tag{7.6}$$

It follows that $\mathcal{K}f(x_i)=g(x_i)$ for all i implies that $\mathcal{K}f(x)=g(x)$ for all $x\in\Sigma$, e.g., if $g\in C(\Sigma)$, and $f\in L^1(\Omega)$. We let $r_n : C(\Sigma)\longrightarrow I\!R^n$ be the restriction operator, defined by

$$r_ng(i)=g(x_i),\quad i=1,2,\ldots,n, \tag{7.7}$$

and let $\pi_n : I\!R^n\longrightarrow C(\Sigma)$ be a prolongation operator, e.g., by means of spline interpolation of some appropriate order. We let $\Pi_n=\pi_nr_n$, and we assume that for all $F\in C(\Sigma)$,

$$\|F-\Pi_nF\|_{C(\Sigma)}\longrightarrow 0\quad\text{as } n\longrightarrow\infty; \tag{7.8}$$

cf. (6.5). This should not be too much of a requirement in view of (7.6). We need the following approximation number for the operator $\mathcal{K}$, viz.

$$\kappa_n=\sup\{\|(I-\Pi_n)\mathcal{K}f\| : \|f\|_{L^1(\Omega)}\leq 1\}, \tag{7.9}$$

which provides a measure of how well the data $[\mathcal{K}f](x_i)$, $i=1,2,\ldots,n$, describe the function $[\mathcal{K}f](x)$, $x\in\Sigma$. By (7.8) we have $\kappa_n\longrightarrow 0$ as $n\longrightarrow\infty$.

THEOREM 7.2. *Let $\varphi\in L^1(\Omega)$ be nonnegative, let $g\in C(\Sigma)$, and $g_i\equiv g(x_i)$ for all i. Suppose that f_o is bounded away from zero, and that f_o solves* (7.4). *Assume that f_o satisfies condition* (4.11). *Let f_n solve* (7.1). *Then $\{f_n\}_n$ converges to the solution of* (7.4), *and for some constant c depending on g and φ,*

$$\|f_n-f_o\|_{L^1(\Omega)}\leq c\kappa_n.$$

Proof. From Corollary 3.2 we get that

$$D(f_o,f_n)\leq D(f_o,\varphi)-D(f_n,\varphi)\leq\int_\Omega(f_o-f_n)\log\frac{f_o}{\varphi}.$$

By the assumption (4.11) the right-hand side may be written as $\int_\Sigma\varrho_o\mathcal{K}(f_o-f_n)$, and since $\Pi_n\mathcal{K}(f_o-f_n)=\Pi_n(g-\mathcal{K}f_n)=0$, this equals

$$\int_\Sigma\varrho_o[(I-\Pi_n)\mathcal{K}(f_n-f_o)]\leq\|\varrho_o\|\cdot\kappa_n\,\|f_n-f_o\|_{L^1(\Omega)}. \tag{7.10}$$

In the proof of the previous theorem (see (7.5)), it was shown that $D(f_n,\varphi)\leq D(f_o,\varphi)$, so that $\{f_n\}_n$ is bounded in $L^1(\Omega)$, see Lemma 2.3. Applying (1.7) to the left-hand side of (7.9) then results in the required estimate. □

Actual estimates follow from the theorem once the precise setting is specified, as in §6.

REFERENCES

[1] U. Amato and W. Hughes, *Maximum entropy regularization of Fredholm integral equations of the first kind*, Inverse Problems, 7 (1991), pp. 793–808.

[2] J. M. Borwein and C. S. Lewis, *Convergence of best entropy estimates*, SIAM J. Optim., 1 (1991), pp. 191–205.

[3] L. M. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Comput. Math. and Math. Phys., 7 (1967), pp. 200–217.

[4] J. P. Butler, J. A. Reeds, and S. V. Dawson, *Estimating solutions of first kind integral equations with nonnegative constraints and optimal smoothing*, SIAM J. Numer. Anal., 18 (1981), pp. 381–397.

[5] Y. Censor, *Row-action methods for huge and sparse systems and their applications*, SIAM Rev., 23 (1981), pp. 444–464.

[6] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM J. Optim., to appear.

[7] I. Csiszár and G. Tusnàdy, *Information geometry and alternating minimization procedures*, Statistics and Decisions, Supplement Issue No. 1 (1984), pp. 205–237.

[8] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[9] H. W. Engl and G. Landl, *Convergence rates for maximum entropy regularization*, SIAM J. Numer. Anal., to appear.

[10] B. R. Frieden, *Restoring with maximum likelihood and maximum entropy*, J. Opt. Soc. Amer., 62 (1972), pp. 511–518.

[11] P. J. Green, *On use of the EM algorithm for penalized likelihood estimation*, J. R. Statist. Soc., B 52 (1990), pp. 443–452.

[12] C. W. Groetsch, *The Theory of Tikhonov Regularization for Fredholm Integral Equations of the First Kind*, Pitman, Boston, 1984.

[13] R. Gordon and G. T. Herman, *Reconstruction of pictures from their projections*, Quart. Bull. Center for Theor. Biology, 4 (1971), pp. 71–151.

[14] R. B. Holmes, *Geometric functional analysis*, Springer-Verlag, New York, 1975.

[15] J. H. B. Kemperman, *On the optimum rate of transmitting information*, Lecture Notes in Math., 23 (1967), pp. 126–169.

[16] M. Klaus and R. T. Smith, *A Hilbert space approach to maximum entropy reconstruction*, Math. Meth. Appl. Sci., 10 (1988), pp. 397–406.

[17] F. M. Larkin, *Estimation of a nonnegative function*, BIT 2 (1969), pp. 30–52.

[18] M. A. Lukas, *Convergence rates for regularized solutions*, Math. Comput., 51 (1988), pp. 107–131.

[19] M. Z. Nashed and G. Wahba, *Generalized inverses in reproducing kernel spaces: An approach to regularization of linear operator equations*, SIAM J. Math. Anal., 5 (1974), pp. 974–987.

[20] F. Natterer, *Error bounds for Tikhonov regularization in Hilbert scales*, Appl. Anal., 18 (1984), pp. 29–37.

[21] D. W. Nychka and D. D. Cox, *Convergence rates for regularized solutions of integral equations from discrete noisy data*, Ann. Statist., 17 (1989), pp. 556–572.

[22] J. Skilling, ed., *Maximum entropy and Bayesian methods*, Kluwer, Dordrecht, 1989.

[23] R. T. Smith and C. K. Zoltani, *An application of the finite element method to maximum entropy tomographic image reconstruction*, J. Sci. Comput., 2 (1987), pp. 283–295.

[24] S. J. Wernecke and L. R. d'Addario, *Maximum entropy image reconstruction*, IEEE Trans. Comput., C-26 (1977), pp. 351–364.

SIAM J. MATH. ANAL.
Vol. 24, No. 6, pp. 1577–1596, November 1993

009

BIFURCATION FROM A PERIODIC ORBIT FOR A STRONGLY RESONANT REVERSIBLE AUTONOMOUS VECTOR FIELD*

MARIE-CHRISTINE PÉROUÈME[†]

Abstract. The author considers a reversible system admitting a symmetric periodic orbit such that the Jordan block belonging to the Floquet exponent zero is four-dimensional, nonsemisimple. Using the normal form theory around closed orbits, it is shown that, generically, such a solution is part of a one-parameter family of symmetric periodic orbits. The existence in such a system of two one-parameter families of symmetric solutions homoclinic to some periodic orbits is also proven. Finally, the author shows how this problem is related to the perturbed reversible 1-1 resonance vector fields, and allows its study to be completed.

Key words. reversible systems, Floquet exponents, normal form theory, integrable systems, homoclinic solutions, 1-1 resonance, Eckhaus instability

AMS subject classifications. 34C20, 23, 25, 37

1. Introduction.

Let us consider a four-dimensional reversible system $du/dt = F(u)$ with a symmetry S such that $F(Su) = -SF(u)$, and let us assume that this system admits a *reversible or symmetric* T-periodic solution $u_0(t)$ such that $Su_0(t) = u_0(-t)$. It is proved in [6] that u_0 belongs to a one-parameter family of periodic solutions, and, therefore, the Floquet exponent zero attached to the solution u_0 is nonsemisimple. In [6], the author studied the case when the Jordan block belonging to the Floquet exponent zero of the operator $du/dt - DF\big(u_0(t)\big)u$ is two-dimensional; because of the reversibility the other exponents are $\pm\lambda$ with $\lambda \neq 0$ generically. He proved the persistence of the family of periodic solutions under small reversible perturbations as well as the existence of branches of subharmonic solutions when $\exp \lambda T$ is a root of unity. (A complete analysis was made in the case when $\exp \lambda T = -1$.)

In the present work we study the case when the Floquet exponent $\lambda \to 0$. More precisely, if we consider a family (parametrized by ν) of reversible systems with a reversible periodic orbit, the reversibility implies that for each value ν this system admits the Floquet exponents zero (nonsemisimple) and $\pm\lambda(\nu)$. Then, we only need ν to be a one-dimensional parameter to impose the value of λ to be zero at a prescribed value, say $\nu = 0$. As two eigenvalues collapse in a nonsimple way generically, the Jordan block belonging to the Floquet exponent zero is then of the type

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

So this singularity of a vector field is a priori of codimension one. Using the normal form theory for flows near closed orbits for an autonomous vector field as developed in [4], we shall give below a proof that it is in fact of codimension zero.

In §2, we compute *a reversible normal form around the periodic orbit.* It turns out that this four-dimensional normal form admits two first integrals and, therefore,

*Received by the editors January 20, 1992; accepted for publication (in revised form) January 5, 1993.

[†] Institut Non Lineaire de Nice, Unité Mixte de Recherche du Centre National de la Recherche Scientifique 129, Université de Nice Sophia-Antipolis, Parc Valrose, F-06108 Nice Cedex 2, France.

is *integrable*. We study extensively the reduced system, focusing on *periodic and homoclinic* solutions.

In §3, we prove that our original system admits in fact a family of symmetric periodic solutions, u_{ψ_0}, which stability changes type [semi-elliptic ↔ semihyperbolic] precisely at the solution u_0. Moreover, we prove that two symmetric homoclinic orbits are associated to each periodic solution of semihyperbolic type.

In §4, we study a small reversible perturbation of our system and prove that generically the solution u_0 is stable. It follows that the families of periodic and homoclinic orbits are also stable and that our singularity was in fact of codimension zero.

Finally, we give an example where such a situation occurs: in [5], studying the reversible 1-1 resonance, the authors proved the existence, in the supercritical case, of a point E, *the Eckhaus point*, where periodic solutions change type (elliptic ↔ hyperbolic). In §5, we prove that at the point E, *the Jordan block belonging to the Floquet exponent zero is four-dimensional.* We then compute the first coefficient of the expansion of the normal form to prove that generically the study of the neighborhood of the Eckhaus point falls within the scope of §§3 and 4. So our Fig. 1 is just the same as Fig. 3 of [5], but here we look at a different scaling. Finally, the present study completes the study of the neighborhood of the point E of [5].

2. The "reversible normal form."

2.1. The derivation of the reduced equation. Consider a four-dimensional differential equation

$$\frac{dx}{dt} = F(x), \tag{2.1}$$

where $x \in \mathbb{R}^4$ and F is a *reversible* vector field. That is, we assume that there exists a symmetry S in $GL(\mathbb{R}^4)$ such that $S^2 = \mathrm{Id}_{\mathbb{R}^4}$, and

$$F(Sx) = -SF(x). \tag{2.2}$$

This implies that if $x(t)$ is a solution of (2.1), so is $\tilde{x}(t) =: Sx(-t)$. A solution such that $x = \tilde{x}$ will be called *a reversible solution.*

Let us make the following assumptions.

(H_1) Equation (2.1) admits a reversible solution, periodic of period T, say $u_0(t)$. We shall denote by

$*L(t) =: D_xF\big(u_0(t)\big)$, the linearized operator around the periodic solution,

$*\mathcal{S}(t)$, the fundamental linear operator, solution of the equation

$$\begin{aligned} \frac{d}{dt}\mathcal{S}(t) &= L(t)\mathcal{S}(t), \\ \mathcal{S}(0) &= Id_{\mathbb{R}^4}, \end{aligned} \tag{2.3}$$

$*\mathcal{S}(T)$, the monodromy operator.

It can be easily proved that, as (2.1) is autonomous, $\dot{u}_0(t) = \mathcal{S}(t)\dot{u}_0(0)$ and that, as a consequence, $\dot{u}_0(0)$ is an eigenvector of $S(T)$ belonging to the eigenvalue 1.

Our next assumption is as follows.

(H_2) The Jordan block belonging to the eigenvector $\dot{u}_0(0)$ is four-dimensional, i.e., as

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

According to [4], we can find a system of independent T-periodic vector-functions satisfying

$$\left(-\frac{d}{dt}+L(t)\right)\zeta_i(t)=\zeta_{i-1}(t), \qquad i=0\cdots 3, \quad \text{with the convention } \zeta_{-1}\equiv 0. \tag{2.4}$$

Note that $\zeta_0(t)=\dot{u}_0(t)$ and satisfies $\tilde{\zeta}_0=-\zeta_0$. It follows that $\tilde{\zeta}_1$ is a solution of the same equation as ζ_1. Then, we have $\tilde{\zeta}_1-\zeta_1=\alpha\zeta_0$ and replacing ζ_1 by $\zeta_1+(\alpha/2)\zeta_0$, we can assume that $\tilde{\zeta}_1=\zeta_1$. This now implies that $-\tilde{\zeta}_2$ is a solution of the same equation as ζ_2, i.e., that $\tilde{\zeta}_2+\zeta_2=\beta\zeta_0$. Applying $\tilde{}$, we can deduce that $\tilde{\zeta}_2+\zeta_2=\beta\tilde{\zeta}_0=-\beta\zeta_0$, i.e., that $\tilde{\zeta}_2=-\zeta_2$. Using the same argument, we finally prove that we can find the ζ_i's such that

$$S\zeta_i(-t)=(-1)^{i+1}\zeta_i(t). \tag{2.5}$$

It is proved in [4] that there exists a nonlinear change of variables (normal form) in the neighborhood of the periodic orbit $\Gamma =: \{u_0(t),\ t\in\mathbb{R}\}$:

$$x(t)=u_0(\tau)+\psi\zeta_1(\tau)+A\zeta_2(\tau)+B\zeta_3(\tau)+\Phi(\tau,y), \tag{2.6}$$

such that *at any arbitrary order* (2.1) becomes an autonomous system:

$$\begin{aligned}\frac{d\tau}{dt} &= 1+\psi+\mathrm{n}\,(y),\\ \frac{dy}{d\tau} &= L_0y+N(y),\end{aligned} \tag{2.7}$$

where

$$y=:\begin{pmatrix}\psi\\ A\\ B\end{pmatrix} \quad\text{and}\quad L_0=:\begin{pmatrix}0&1&0\\0&0&1\\0&0&0\end{pmatrix},$$

Φ is T-periodic and smooth in y, n is a polynomial, N is a polynomial vector function, Φ, n, and N are $O(|y|^2)$.

The normal form theory developed in [4] shows that n and N have to commute with $e^{\tau L_0^*}$; hence, it is proved in [2] that a good choice of n and N is as follows:

$$\begin{aligned}\mathrm{n}\,(y)&=n(\psi,A^2-2\psi B),\\ N(y)&=\begin{pmatrix}\psi P_1(\psi,A^2-2\psi B)\\ AP_1(\psi,A^2-2\psi B)+\psi P_2(\psi,A^2-2\psi B)\\ BP_1(\psi,A^2-2\psi B)+AP_2(\psi,A^2-2\psi B)+\psi P_3(\psi,A^2-2\psi B)\end{pmatrix},\end{aligned}$$

where P_j are polynomial in their arguments. In our case, as the system is reversible and as the ζ_i's satisfy (2.5), Φ can be choosen so that it respects the symmetry:

$$S':\begin{pmatrix}\psi\\ A\\ B\end{pmatrix}\mapsto\begin{pmatrix}\psi\\ -A\\ B\end{pmatrix}, \tag{2.8}$$

and N can be chosen so that the new system in (y,τ) is reversible with respect to S', i.e.,

$$\Phi(-\tau,S'y)=S\Phi(\tau,y), \qquad N(S'y)=-S'N(y), \tag{2.9}$$

which implies that $P_1 \equiv P_3 \equiv 0$.

Finally, up to *any arbitrary order* (2.1) can be written as

$$\tag{2.10}\begin{cases} \dfrac{d\tau}{dt} = 1 + \psi + n(\psi, A^2 - 2\psi B), \\ \dfrac{d\psi}{d\tau} = A, \\ \dfrac{dA}{d\tau} = B + \psi P(\psi, A^2 - 2\psi B), \\ \dfrac{dB}{d\tau} = AP(\psi, A^2 - 2\psi B). \end{cases}$$

By the expression *at an arbitrary order* we mean that (2.1) writes

$$\tag{2.11}\begin{cases} \dfrac{d\tau}{dt} = 1 + \psi + n(\psi, A^2 - 2\psi B) + \text{h.o.t}, \\ \dfrac{d\psi}{d\tau} = A + \text{h.o.t}, \\ \dfrac{dA}{d\tau} = B + \psi P(\psi, A^2 - 2\psi B) + \text{h.o.t}, \\ \dfrac{dB}{d\tau} = AP(\psi, A^2 - 2\psi B) + \text{h.o.t}, \end{cases}$$

where h.o.t stands for T-periodic functions which are $O(\|y\|^{N+1})$, with an arbitrary N.

The solution $u_0(t)$ of (2.1) now corresponds to the solution $y = 0$ of (2.10), and the linearized equation around the basic periodic solution for the transverse dynamics is

$$\tag{2.12}\frac{dy}{d\tau} = L_0 y.$$

We can sum up these results saying that we have found *a reversible normal form for* (2.1) *around the periodic orbit* Γ. For more results about reduction of vector fields with a periodic orbit, see [1].

2.2. Properties of the reduced equations. It is easy to check that (2.10) admits two first integrals:

$$\tag{2.13}\begin{cases} K = A^2 - 2\psi B, \\ H = B - G(\psi, K), \end{cases} \quad \text{where } G(\psi, K) =: \int_0^\psi P(s, K)ds.$$

As a consequence, (2.10) is integrable and can be written as

$$\tag{2.14}\begin{cases} \left(\dfrac{d\psi}{d\tau}\right)^2 = 2\psi\big(G(\psi, K) + H\big) + K, \\ \dfrac{dA}{d\tau} = G(\psi, K) + H + \psi P(\psi, K), \\ B = G(\psi, K) + H, \\ \dfrac{d\tau}{dt} = 1 + \psi + n\big(\psi, K\big). \end{cases}$$

Let us define $f(\psi, K, H) =: 2\psi\bigl(G(\psi, K) + H\bigr) + K$, in such a way that the equations satisfied by y are

$$\left\{\begin{aligned} \frac{d\psi}{d\tau} &= \varepsilon\sqrt{f(\psi, K, H)}, \\ \frac{dA}{d\tau} &= f'_\psi(\psi, K, H)/2, \qquad\qquad \text{with } \varepsilon = \operatorname{sgn} A. \\ \frac{dB}{d\tau} &= \varepsilon\sqrt{f(\psi, K, H)}P(\psi, K), \end{aligned}\right. \tag{2.15}$$

We can notice that fixed points of (2.15) are of the type

$$y = \begin{pmatrix} \psi_0 \\ 0 \\ B \end{pmatrix},$$

where (i) ψ_0 is a double root of the polynomial f, i.e., satisfies

$$\left\{\begin{aligned} 2\psi_0\bigl(G(\psi_0, K) + H\bigr) + K = 0, \\ G(\psi_0, K) + H + \psi_0 P(\psi_0, K) = 0, \end{aligned}\right. \tag{2.16}$$

and (ii) taking account of $A = dA/d\tau = 0$, B satisfies

$$B + \psi_0 P(\psi_0, -2\psi_0 B) = 0. \tag{2.17}$$

For ψ_0 small enough, it is possible to solve (2.17) with respect to B using the implicit function theorem. And finally, fixed points of (2.15) write

$$y = \begin{pmatrix} \psi_0 \\ 0 \\ B(\psi_0) \end{pmatrix},$$

where ψ_0 satisfies (2.16) and $B(\psi_0)$ (2.17). Now, letting ψ_0 be a double root of f, the linearized operator in y around the corresponding fixed point is

$$L(\psi_0) = \begin{pmatrix} 0 & 1 & 0 \\ P_0 + \psi_0 P'_{\psi_0} - 2\psi_0 B(\psi_0) P'_K & 0 & 1 - 2\psi_0{}^2 P'_K \\ 0 & P_0 & 0 \end{pmatrix}, \tag{2.18}$$

where $P_0 = P\bigl(\psi_0, -2\psi_0 B(\psi_0)\bigr)$. Its eigenvalues are zero, due to the existence of a family of fixed points, and $\pm\lambda$, where

$$\lambda^2 = 2P_0 + \psi_0 P'_{\psi_0}. \tag{2.19}$$

2.3. Fixed points and homoclinic solutions of the reduced equations in the generic case. Let us present the expansions we use below:

$$P(\psi, K) = \alpha\psi + \beta K + \cdots, \tag{2.20}$$

$$f(\nu\psi, K, H) = \alpha\psi^3 + 2\beta K\psi^2 + 2H\psi + K + \cdots. \tag{2.21}$$

By "generic case" we mean the case $\alpha \neq 0$. The principal part, for small ψ_0, of the set where we find double roots for f is then given by (see Fig. 1)

$$\left\{\begin{aligned} K &= 2\alpha\psi_0{}^3, \\ H &= -3\alpha\psi_0{}^2/2, \end{aligned}\right. \tag{2.22}$$

and

$$\lambda^2 = 3\alpha\psi_0 + O(\psi_0{}^2). \tag{2.23}$$

The origin is a singular point, corresponding to a triple root of f, where all the eigenvalues vanish in a nonsimple way. The Jordan block corresponding to the eigenvalue zero is then three-dimensional.

If we complete our figure, adding the form of the graph of f and the eigenvalues of $L(\psi_0)$, it appears that there exists a one-parameter family of hyperbolic fixed points along the arc OG. This family can be parametrized by $\psi_0 > 0$, and we shall denote it by

$$y_{\psi_0} =: \begin{pmatrix} \psi_0 \\ 0 \\ B(\psi_0) \end{pmatrix}. \tag{2.24}$$

Let us also denote by $H(\psi_0), K(\psi_0)$, and λ_{ψ_0} the corresponding values of H, K, and λ. For each $\psi_0 > 0$ there exist orbits homoclinic to y_{ψ_0} which are of the type

$$h_{\psi_0}(\tau,\tau_0) =: \begin{pmatrix} \psi_0(\tau-\tau_0) \\ A(\psi_0(\tau-\tau_0)) \\ B(\psi_0(\tau-\tau_0)) \end{pmatrix} \quad \underset{\text{as } \tau\ \rightarrow\ \pm\infty}{\longrightarrow} \quad \begin{pmatrix} \psi_0 \\ 0 \\ B(\psi_0) \end{pmatrix}, \tag{2.25}$$

where (i) the function $\psi_0(\tau)$ is defined implicitly by the formula

$$|\tau| = \operatorname{sgn}\alpha \int_{\psi_0{}'}^{\psi_0(\tau)} \frac{ds}{\sqrt{f(s, K(\psi_0), H(\psi_0))}}, \tag{2.26}$$

(ii) $\psi_0{}'$ is the other root of f which is simple,

(iii) $A(\psi) = \operatorname{sgn}(\alpha(\psi - \psi_0{}'))\sqrt{f(\psi, K(\psi_0), H(\psi_0))}$, $B(\psi) = G(\psi, K(\psi_0)) + H(\psi_0)$.

It follows from the definition that $\psi_0(.)$ is an even function of τ and that $\psi_0(0) = \psi_0{}'$. As a consequence, $A(\psi_0(0)) = 0$, and among the homoclinic orbits one is reversible: the one corresponding to the choice $\tau_0 = 0$. We shall denote it by $h_{\psi_0}(\tau)$.

3. Solution homoclinic to a periodic orbit in the generic case. In the same way as (2.10) leads to (2.15), which can be written as

$$\frac{dy}{d\tau} = F'(y), \tag{3.1}$$

(2.11) leads to

$$\frac{dy}{d\tau} = F'(y) + G'(y,\tau), \quad \text{with } G'(y,\tau) = O(\|y\|^{N+1}). \tag{3.2}$$

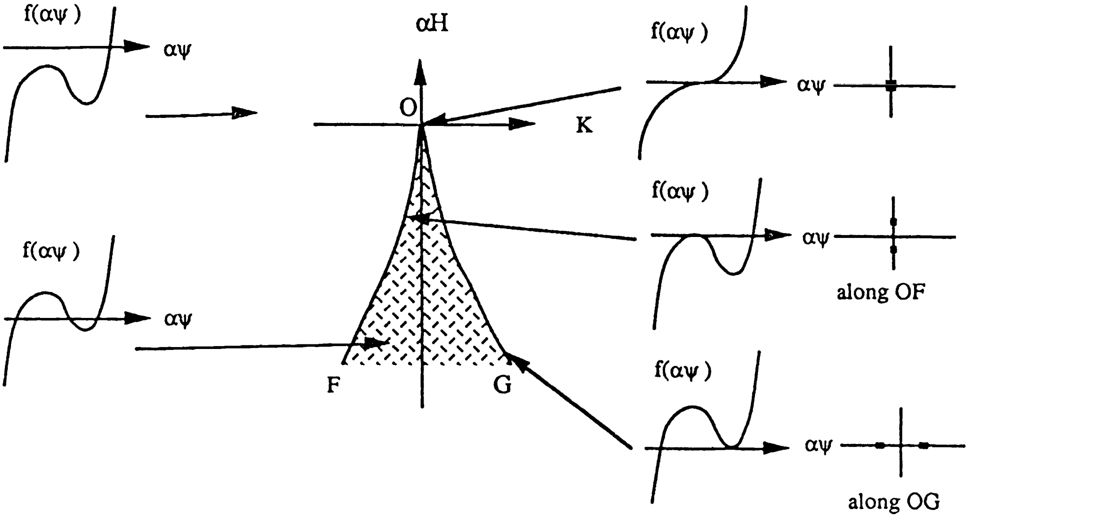


FIG. 1. *Different shapes of the graph of the function f and Floquet exponents of periodic orbits.*

Moreover, G' is T-periodic and reversible:

$$G'(S'y,-\tau) = -S'G'(y,\tau). \tag{3.3}$$

In the previous section, we proved that (3.1) admits reversible fixed points and homoclinic orbits. Our aim is now to prove their persistence for (3.2).

Each solution $y(\tau)$ of (3.2) leads to a one-parameter family of solutions of (2.1):

$$\begin{aligned} x(t) &= u_0(\tau) + \psi(\tau)\zeta_1(\tau) + A(\tau)\zeta_2(\tau) + B(\tau)\zeta_3(\tau) + \Phi(\tau, y(\tau)), \\ t &= \int_{\tau_0}^{\tau} \frac{d\tau}{1+\psi(\tau)+n(\psi(\tau),K)}. \end{aligned} \tag{3.4}$$

Note that $S\zeta_i(T/2) = (-1)^{i+1}\zeta_i(T/2)$. So, if $y(\tau)$ is a reversible solution of (3.2) among the corresponding family of solutions of (2.1) two are reversible: the one corresponding to the choice $\tau_0 = 0$ or $\tau_0 = T/2$.

3.1. Existence of periodic solutions. The proof of the existence of reversible periodic orbits of (3.2) will go in two steps.

The first step consists in controlling the size of the perturbed solutions close to some fixed point of (3.1). We shall follow in that part the proof of [5], that is, we shall show that such perturbed orbits are solutions of a functional equation, and solving this equation will give us, at the same time, the size of these solutions.

The second step will look for periodic reversible solutions. Our approach here will be different from the one of [5]: we shall not involve geometrical arguments but we shall show that the previous functional equation admits a solution in the space of periodic reversible functions.

Let us look for a solution of (3.2) of the form $y(\tau) = y_{\psi_0} + z(\tau)$, $\|z\| \ll \|y_{\psi_0}\| = O(\psi_0)$. Note that ψ_0 appears here as an additional free parameter. This allows us to choose $z(0)$ in any convenient direction transverse to $dy_{\psi_0}/d\psi_0$. And z has to satisfy

$$\frac{dz}{d\tau} - L(\psi_0)z = N(z) + R(z,\tau), \tag{3.5}$$

where $N(z) = F(y_{\psi_0} + z) - F(y_{\psi_0}) - L(\psi_0)z$, and $R(z,\tau) = G(y_{\psi_0} + z, \tau)$, or equivalently,

$$z(\tau) = e^{\tau L(\psi_0)}\xi + \int_0^\tau e^{(\tau - s)L(\psi_0)}\Big(N\big(z(s)\big) + R\big(z(s), s\big)\Big)ds, \tag{3.6}$$

with $\xi = z(0)$. We can notice that

$$\begin{aligned} N(z) &= O(\|z\|^2), \\ N(z) - N(z') &= O\big((\|z\| + \|z'\|) \cdot \|z - z'\|\big), \\ R(z,\tau) &= O({\psi_0}^{N+1}), \\ R(z,\tau) - R(z',\tau) &= O({\psi_0}^N \|z - z'\|). \end{aligned} \tag{3.7}$$

Then, if we denote by Θ_ξ the right-hand side of (3.6), and if we let z and z' in $\mathcal{C}_d$, the set of continuous functions $[-T/2, T/2] \rightarrow \mathrm{IR}^4$ such that $\|z\|_T =: \sup_{\tau \in [-T/2,T/2]} \|z(\tau)\| \leq d$, there exists M such that

$$\begin{aligned} \|\Theta_\xi z\|_T &\leq M\left(\|\xi\| + (d^2 + {\psi_0}^{N+1})\right), \\ \|\Theta_\xi z - \Theta_\xi z'\|_T &\leq M(d + {\psi_0}^N)\ \|z - z'\|. \end{aligned} \tag{3.8}$$

We see that one can always find ψ_0, d, and $\|\xi\|$ small enough for Θ_ξ to be a contraction on $\mathcal{C}_d$. Now, it is a classical result that, when Θ_ξ is a contraction, (3.6) admits an unique solution in $\mathcal{C}_d$, say z_ξ.

Moreover, due to the reversibility, $S'z_\xi(-t)$ is another solution of (3.6) starting at $S'\xi$ at $t = 0$, i.e.,

$$S'z_\xi(-t) = z_{S'\xi}(t).$$

It follows that if we choose $\xi \in \Pi =: \{y/\ S'y = y\}$, the corresponding z_ξ is reversible. And z_ξ will lead to a T-periodic solution of (3.2) if and only if it satisfies the additional condition $z_\xi(-T/2) = z_\xi(T/2) = S'z_\xi(T/2)$, i.e., if and only if

$$\begin{aligned} \Lambda_{\psi_0}\xi = &- \int_0^{T/2} e^{(T/2 - s)L(\psi_0)}\Big(N\big(z_\xi(s)\big) + R\big(z_\xi(s), s\big)\Big)ds \\ &- \int_{-T/2}^{0} e^{-(T/2 + s)L(\psi_0)}\Big(N\big(z_\xi(s)\big) + R\big(z_\xi(s), s\big)\Big)ds, \end{aligned} \tag{3.9}$$

where $\Lambda_{\psi_0} =: e^{(T/2)L(\psi_0)} - e^{-(T/2)L(\psi_0)}$. Notice that Λ_{ψ_0} is a linear operator mapping Π onto the one-dimensional subspace $\Delta =: \{y/\ S'y = -y\}$, and which nullspace is generated by $dy_{\psi_0}/d\psi_0$. Moreover, one can easily show that $S'L(\psi_0) = -L(\psi_0)S'$ and deduce that

$$S'e^{\tau L(\psi_0)} = e^{-\tau L(\psi_0)}S'.$$

So, for each $\xi \in \Pi$,

$$\begin{aligned}\int_{-T/2}^{0} & e^{-(T/2+s)L(\psi_0)}\Big(N\big(z_\xi(s)\big) + R\big(z_\xi(s), s\big)\Big)ds \\ &= -\int_{-T/2}^{0} S' e^{(T/2+s)L(\psi_0)}\Big(N\big(S' z_\xi(s)\big) + R\big(S' z_\xi(s), -s\big)\Big)ds \\ &= -S' \int_{0}^{T/2} e^{(T/2-s)L(\psi_0)}\Big(N\big(z_\xi(s)\big) + R\big(z_\xi(s), s\big)\Big)ds.\end{aligned}$$

It follows that the right-hand side of (3.9) belongs to Δ and (3.9) reduces in fact to a scalar equation. More precisely, let us denote by $\xi_{\lambda_{\psi_0}}$ and $\xi_{-\lambda_{\psi_0}} =: S'\xi_{\lambda_{\psi_0}}$ the eigenvectors of $L(\psi_0)$ belonging to the eigenvalues $\pm\lambda_{\psi_0}$. Then, $\xi \in \Pi$ writes

$$\xi = \mathsf{a}\frac{dy_{\psi_0}}{d\psi_0} + \mathsf{b}\left(\xi_{\lambda_{\psi_0}} + \xi_{-\lambda_{\psi_0}}\right).$$

As a consequence of our remark on the choice of $z(0)$, we can impose the transversality condition $\mathsf{a} = 0$, and (3.9) can be reformulated as

$$g(\mathsf{b}, \psi_0) = 0, \quad \text{with } g(0,0) = 0. \tag{3.10}$$

Noting that $e^{\tau L(\psi_0)}\xi_{\lambda_{\psi_0}} = e^{\lambda_{\psi_0}\tau}\xi_{\lambda_{\psi_0}}$ and $e^{\tau L(\psi_0)}\xi_{-\lambda_{\psi_0}} = e^{-\lambda_{\psi_0}\tau}\xi_{-\lambda_{\psi_0}}$, one can compute

$$\Lambda_{\psi_0}(\xi_{\lambda_{\psi_0}} + \xi_{-\lambda_{\psi_0}}) = 2\sinh\frac{T\lambda_{\psi_0}}{2}(\xi_{\lambda_{\psi_0}} - \xi_{-\lambda_{\psi_0}}).$$

So, applying (3.7) once again, (3.9) leads to

$$\mathsf{b}\sinh\frac{T\lambda_{\psi_0}}{2} + O(\mathsf{b}^2 + {\psi_0}^{N+1}) = 0, \tag{3.11}$$

or equivalently,

$$\mathsf{b} + O\left(\frac{\mathsf{b}^2 + {\psi_0}^{N+1}}{\lambda_{\psi_0}}\right) = 0. \tag{3.12}$$

This equation can be solved with respect to b provided that ${\psi_0}^{N+1}/\lambda_{\psi_0}^2$ is small enough. We then have $\|\xi\| = \mathsf{b} = O({\psi_0}^{N+1}/\lambda_{\psi_0})$. From (2.23), $\lambda_{\psi_0} = O({\psi_0}^{1/2})$, and, therefore, $\|\xi\| = O({\psi_0}^{2N+1/2})$. Now coming back to (3.8), provided that ${\psi_0}^{2N+1/2} \ll 1$, one can choose $d = O({\psi_0}^{2N+1/2})$. It means that if $N \geq 2$, for each value of $\psi_0 > 0$, small enough, (3.2) admits a reversible periodic solution, say $y_{\psi_0}(\tau)$. Remark that the bigger N is, the smaller one can choose d. This means that increasing the value of N, one improves the distance between the periodic solutions of (3.1) and (3.2).

Remark. Notice that the same analysis applies to prove the persistence of the family of fixed points of (2.1) occuring along the curve OF. The eigenvalues are then: zero, double, and two pure imaginary complex numbers, $\pm\iota\lambda'_{\psi_0}$ $[\psi_0 < 0, \lambda_{\psi_0} \in \mathbb{R}]$. All the calculations are the same except that we now choose $-\iota(\xi_{\lambda'_{\psi_0}} + \xi_{-\lambda'_{\psi_0}})$ as basis vector instead of $\xi_{\lambda_{\psi_0}} + \xi_{-\lambda_{\psi_0}}$, and we have to replace $\sinh(T\lambda_{\psi_0}2)$ by $\sinh(\iota T\lambda'_{\psi_0}/2) = -\sin(T\lambda'_{\psi_0}/2)$.

THEOREM. *The original solution u_0 is in fact part of a family of reversible periodic solutions, say u_{ψ_0}, which are of semi-elliptic type for $\psi_0 < 0$ and of semihyperbolic type for $\psi_0 > 0$.*

3.2. Existence of reversible homoclinic solutions. Here again, we shall follow the proof of [5]. However, in putting our system into a normal form, we have factored out the phase in our equations, we then have a quite simple way to express that we are looking for reversible homoclinic solutions.

Let us look for a solution of (3.2) of the form $y(\tau) = h_{\psi_0}(\tau) - y_{\psi_0} + y_{\psi_0}(\tau) + z(\tau)$. Then z has to satisfy

$$\frac{dz}{d\tau} - DF\big(h_{\psi_0}(\tau)\big)z = N'(z,\tau) + R'(z,\tau), \tag{3.13}$$

where

$$\begin{aligned}N'(z,\tau) =: &F\big(h_{\psi_0}(\tau) - y_{\psi_0} + y_{\psi_0}(\tau) + z(\tau)\big)\\ &- F\big(h_{\psi_0}(\tau)\big) - F\big(y_{\psi_0}(\tau)\big) - DF\big(h_{\psi_0}(\tau)\big)z,\end{aligned}$$

and

$$R'(z,\tau) =: G\big(h_{\psi_0}(\tau) - y_{\psi_0} + y_{\psi_0}(\tau) + z(\tau),\tau\big) - G\big(y_{\psi_0}(\tau),\tau\big).$$

We can prove estimations similar to estimations (3.7):

$$\begin{aligned}\|N'(z,\tau)\| &= O\left(\|z\|^2 + d\,\|z\| + d\,\|h_{\psi_0} - y_{\psi_0}\|\right),\\ \|N'(z,\tau) - N'(z',\tau)\| &= O\big((\|z\| + \|z'\| + d)\,\|z - z'\|\big),\\ \|R'(z,\tau)\| &= O\big({\psi_0}^{N+1}\,(\|z\| + \|h_{\psi_0} - y_{\psi_0}\|)\big),\\ \|R'(z,\tau) - R'(z',\tau)\| &= O\big({\psi_0}^{N}\,\|z - z'\|\big),\end{aligned} \tag{3.14}$$

where d is such that $\|y_{\psi_0}(\tau) - y_{\psi_0}\| \leq d$ and is assumed to be $\ll \psi_0$.

Notice that if z is a solution of (3.13) going to zero as $\tau \to \pm\infty$, the corresponding y a homoclinic solution of (3.2), we look for solutions of (3.11) in

$$E_\mu =: \left\{ z \Big/ \sup_{\tau\,\in\,\mathbb{R}} \|z(\tau)\| e^{\mu|\tau|} < +\infty \right\}, \qquad 0 < \mu < \lambda_{\psi_0},\ \mu \text{ close to } \lambda_{\psi_0}.$$

In order to solve (3.13), we need some information about the linearized equation around $h_{\psi_0}(\tau)$ which reads

$$\frac{dz}{d\tau} - DF\big(h_{\psi_0}(\tau)\big)z = 0. \tag{3.15}$$

Let $p_{\psi_0}(\tau) =: (\partial h_{\psi_0}/\partial\tau)(\tau)$, and $r_{\psi_0}(\tau) =: (\partial h_{\psi_0}/\partial\psi_0)(\tau)$, then one can easily check that both p_{ψ_0} and r_{ψ_0} are solutions of (3.15), such that

$$\begin{aligned}&\|p_{\psi_0}(\tau)\| \simeq e^{-\lambda_{\psi_0}\tau} \quad \text{as} \quad \tau \to +\infty, \quad S'p_{\psi_0}(-\tau) = -p_{\psi_0}(\tau),\\ &r_{\psi_0}(\tau) \to \frac{dy_{\psi_0}}{d\psi_0} \quad \text{as} \quad \tau \to +\infty, \quad S'r_{\psi_0}(-\tau) = r_{\psi_0}(\tau).\end{aligned}$$

Now, looking for instance at the Wronskian, one can see that (3.15) admits a solution $q_{\psi_0}(\tau)$ such that

$$S'q_{\psi_0}(-\tau) = q_{\psi_0}(\tau), \qquad \|q_{\psi_0}(\tau)\| \simeq e^{\lambda_{\psi_0}\tau} \quad \text{as} \quad \tau \to +\infty.$$

Let us denote by $(p^*_{\psi_0}, q^*_{\psi_0}, r^*_{\psi_0})$ the adjoint basis of $(p_{\psi_0}, q_{\psi_0}, r_{\psi_0})$. Then one can check that if z is a solution of the functional equation

$$\begin{aligned}z(\tau) = &\int_0^\tau \langle N'(z(s),s) + R'(z(s),s) \mid p^*_{\psi_0}(s)\rangle ds\; p_{\psi_0}(\tau)\\ &- \int_\tau^{+\infty} \langle N'(z(s),s) + R'(z(s),s) \mid q^*_{\psi_0}(s)\rangle ds\; q_{\psi_0}(\tau)\\ &- \int_\tau^{+\infty} \langle N'(z(s),s) + R'(z(s),s) \mid r^*_{\psi_0}(s)\rangle ds\; r_{\psi_0}(\tau),\end{aligned} \tag{3.16}$$

z is a solution of (3.13) lying in Π at $\tau = 0$, therefore a reversible solution. If we denote by Θ' the right-hand side of (3.16), and by $\|z\|_\mu =: \sup_{\tau \in \mathbb{R}} \|z(\tau)\| e^{\mu|\tau|}$ for $z \in E_\mu$, we now obtain, as for (3.8),

$$\begin{aligned}\|\Theta' z\|_\mu &\le \frac{M'}{\lambda_{\psi_0}} \left(\|z\|^2_\mu + (d + {\psi_0}^{N+1}) \frac{\lambda_{\psi_0}}{\lambda_{\psi_0} - \mu} \right)\\ \|\Theta' z - \Theta' z'\|_\mu &\le \frac{M'}{\lambda_{\psi_0}} \left(\|z\|_\mu + \|z'\|_\mu + (d + {\psi_0}^{N}) \frac{\lambda_{\psi_0}}{\lambda_{\psi_0} - \mu} \right) \|z - z'\|_\mu.\end{aligned} \tag{3.17}$$

We now recall that $\lambda_{\psi_0} = O({\psi_0}^{1/2})$ and that $d = O({\psi_0}^{2N+1/2})$, according to §3. Let us also set $\mu = \lambda_{\psi_0}(1-\epsilon)$. Then, provided that ${\psi_0}^{2N-1/2} \ll \epsilon$, we can find d' such that Θ' is a contraction on the ball of radius d' of E_μ. For instance, one can choose $d' = O({\psi_0}^N/\epsilon)$. So, if $N \ge 2$ and $\psi_0 > 0$ is small enough, (3.2) admits a reversible solution homoclinic to $y_{\psi_0}(\tau)$ and (2.1) admits two reversible solutions homoclinic to $u_{\psi_0}(t)$. Notice that the bigger N is, the smaller one can choose ϵ; it means that increasing the value of N, we can improve the exponential tendency of these homoclinic solutions towards the periodic orbit.

THEOREM. *Equation* (2.1) *admits two reversible solutions homoclinic to each periodic orbit of semihyperbolic type.*

4. Roughness.

We are now interested in proving that these families of periodic and homoclinic solutions of (2.1) are generically stable under small reversible perturbations. From the results of §3 it is enough to prove the persistence of a reversible periodic solution such that the Jordan block belonging to the Floquet exponent zero is of the same type as the one attached to u_0. The method we shall use derives from the implicit function theorem and will give at the same time a shorter proof of the existence of a family of periodic solutions of (2.1). But this method does not allow us to reach homoclinic orbits. Anyway, as we pointed out, the part in §3 where we compute explicitly the distance between periodic orbits of (3.1) and (3.2) is essential in proving the persistence of homoclinic orbits. This is the reason why we include the two proofs.

Let us consider a reversible perturbation of (2.1):

$$\frac{dx}{dt} = F(\nu, x), \qquad F(0,x) = F(x), \quad F(\nu, Sx) = -SF(\nu, x). \tag{4.1}$$

We wonder on which conditions (4.1) admits periodic reversible solutions of period close to T, say $T/(1+\psi)$ with $\psi \simeq 0$. Let us make the change on the time $\tau = (1+\psi)t$, so that we now look for T-periodic reversible solutions of

$$\frac{dx}{d\tau} = \frac{F(\nu, x)}{1+\psi}. \tag{4.2}$$

We can look for solutions of (4.2) of the form $x = u_0 + z$. Then z has to satisfy

$$\begin{aligned}\frac{dz}{d\tau} - D_xF(u_0)z &= \frac{F(\nu, u_0+z)}{1+\psi} - F(u_0) - D_xF(u_0)z \\ &= \nu D_\nu F(0,u_0) - \psi\zeta_0 + o(\|z\| + \nu + \psi).\end{aligned} \tag{4.3}$$

To solve this equation we need some information about the linearized operator

$$\mathcal{A} : z \mapsto -\frac{dz}{d\tau} + D_xF(u_0)z. \tag{4.4}$$

PROPOSITION. *Let C^0_T (respectively, C^1_T) the space of continuous (respectively, continuously derivable) vector T-periodic functions $\mathbb{R} \to \mathbb{R}^4$. Then $\mathcal{A} : C^1_T \to C^0_T$ is a Fredholm operator of index zero. His kernel is generated by ζ_0, and his image is the orthogonal of the space generated by ζ_3^* for the standard scalar product on C^0_T:*

$$\langle z \mid z'\rangle_T =: \frac{1}{T}\int_0^T \langle z(t) \mid z'(t)\rangle\, dt.$$

It follows that if we denote by C^+ (respectively, C^-) the space of vector functions such that $Sz(t) = z(-t)$ (respectively, $Sz(t) = -z(-t)$), $\mathcal{A}$ is an inversible operator $C^1_T{}^+ \to C^0_T{}^-$.

Proof. Let $\big(\zeta_0^*(t), \zeta_1^*(t), \zeta_2^*(t), \zeta_3^*(t)\big)$ be the adjoint basis of $\big(\zeta_0(t), \zeta_1(t), \zeta_2(t), \zeta_3(t)\big)$ for the standard scalar product on $\mathbb{R}^4$, and ${}^t\mathcal{A}$ the adjoint operator of $\mathcal{A}$:

$${}^t\mathcal{A} : z \mapsto \frac{dz}{d\tau} + {}^tD_xF(u_0)z. \tag{4.5}$$

Then one can easily check that

$${}^t\mathcal{A}\zeta_i^* = \zeta_{i+1}^*, \qquad i = 0\cdots 3, \quad \text{with the convention } \zeta_4^* \equiv 0. \tag{4.6}$$

Moreover, let $z \in C^1$ and $f \in C^0_T$. Then

$$\mathcal{A}z = f \text{ and } z \in C^1_T \text{ iff } \quad \begin{array}{lclcl} \dfrac{d\langle z \mid \zeta_3^*\rangle}{d\tau} &=& -\langle f \mid \zeta_3^*\rangle & \text{and} & \int_0^T \dfrac{d\langle z \mid \zeta_3^*\rangle}{d\tau} = 0, \\ \dfrac{d\langle z \mid \zeta_2^*\rangle}{d\tau} &=& -\langle f \mid \zeta_2^*\rangle + \langle z \mid \zeta_3^*\rangle & \text{and} & \int_0^T \dfrac{d\langle z \mid \zeta_2^*\rangle}{d\tau} = 0, \\ \dfrac{d\langle z \mid \zeta_1^*\rangle}{d\tau} &=& -\langle f \mid \zeta_1^*\rangle + \langle z \mid \zeta_2^*\rangle & \text{and} & \int_0^T \dfrac{d\langle z \mid \zeta_1^*\rangle}{d\tau} = 0, \\ \dfrac{d\langle z \mid \zeta_0^*\rangle}{d\tau} &=& -\langle f \mid \zeta_0^*\rangle + \langle z \mid \zeta_1^*\rangle & \text{and} & \int_0^T \dfrac{d\langle z \mid \zeta_0^*\rangle}{d\tau} = 0. \end{array} \tag{4.7}$$

It is easy to see that provided that $\langle f \mid \zeta_3^*\rangle_T = 0$, these equations admit a solution which is unique up to adding a multiple of ζ_0 and that this solution satisfies

$$\langle z \mid \zeta_{i+1}^*\rangle_T = \langle f \mid \zeta_i^*\rangle_T, \qquad i = 0, 1, 2. \tag{4.8}$$

This proves the first part of the proposition. The other is a consequence of the fact that $\mathcal{A}$ maps C^{1-} into C^{0+}, C^{1+} into C^{0-}, and $\zeta_0 \in C^{1-}, \zeta_3^* \in C^{1+}$.

Then, using the implicit function theorem, (4.3) can be solved with respect to $z \in \mathcal{C}_T^{1+}$ provided that ν and ψ are small enough. Let us denote by $z(\nu, \psi, \tau)$ this solution. The linearized equation around the corresponding T-periodic orbit of (4.3) writes

$$\frac{dz}{d\tau} = D_xF(\nu, u_0 + z(\nu, \psi))z/(1+\psi). \tag{4.9}$$

We shall denote by $\mathcal{S}(\nu, \psi, \tau)$ its fundamental matrix. Due to the existence of a family of periodic solutions, 1 is a Floquet multiplier of $\mathcal{S}(\nu, \psi, T)$ at least double nonsemisimple. Moreover, one can easily check that

$$\det \mathcal{S}(\nu, \psi, T) = \exp \int_0^T \operatorname{trace} D_xF(\nu, u_0 + z(\nu, \psi))/(1+\psi),$$

and that, due to the reversibility, $\det \mathcal{S}(\nu, \psi, T) = 1$. So, the other Floquet multipliers are, a priori, $\exp \pm T\lambda_{\nu,\psi}$, with $\lambda_{\nu,\psi} \in \mathbb{R}$ or $\iota\mathbb{R}$, and $\operatorname{trace} \mathcal{S}(\nu, \psi, T) = 2(1 + \cosh(T\lambda_{\nu,\psi}))$.

(All assertions concerning Floquet theory are proved in [3], and the way they are modified by reversibility is extensively studied in [6].)

For a given value of (ν, ψ) all the Floquet multipliers are 1 if and only if

$$\operatorname{trace} \mathcal{S}(\nu, \psi, T) = 4. \tag{4.10}$$

We then have $(\mathcal{S}(\nu, \psi, T) - \mathrm{Id}_{\mathbb{R}^4})^4 = 0_{\mathbb{R}^4}$, and, as $(\mathcal{S}(T) - \mathrm{Id}_{\mathbb{R}^4})^3 \neq 0_{\mathbb{R}^4}$, we still have $(\mathcal{S}(\nu, \psi, T) - \mathrm{Id}_{\mathbb{R}^4})^3 \neq 0_{\mathbb{R}^4}$. It follows that if (4.10) is satisfied the Jordan block belonging to the Floquet multiplier 1 is of the same type as the one attached to u_0. Now equation (4.10) can be solved with respect to ψ provided that

$$\operatorname{trace} \frac{\partial \mathcal{S}}{\partial \psi}(0, 0, T) \neq 0. \tag{4.11}$$

For the normal form, the Floquet exponents are computed in (2.23), and

$$\begin{aligned} \operatorname{trace} \mathcal{S}(\psi_0, T) &= 2(1 + \cosh(T\lambda_{\psi_0})), \\ \operatorname{trace} \frac{\partial \mathcal{S}}{\partial \psi_0}(\psi_0, T) &= 2T \sinh(T\lambda_{\psi_0}) \frac{d\lambda_{\psi_0}}{d\psi_0}. \end{aligned}$$

From (2.23), $2\lambda d\lambda/d\psi_0 \to 3\alpha$ as $\psi_0 \to 0$, and we can deduce that

$$\operatorname{trace} \frac{\partial \mathcal{S}}{\partial \psi_0}(0, T) = 3\alpha T^2. \tag{4.12}$$

We are now going to prove that (4.12) still holds for the complete equation (4.1).

Note that, from (4.3), $\partial z/\partial \psi(0, 0, \tau) \in \mathcal{C}_T^{1+}$ and is a solution of

$$-\frac{dz}{d\tau} + D_xF(u_0)z = \zeta_0.$$

It follows that

$$\frac{\partial z}{\partial \psi}(0, 0, \tau) = \zeta_1(\tau). \tag{4.13}$$

In the same way, from (4.9) and (4.13), $\partial\mathcal{S}/\partial\psi$ is a solution of

$$-\frac{dZ}{d\tau}+D_xF(u_0)Z=\big(D_xF(u_0)-D_{xx}F(u_0)\zeta_1\big)\mathcal{S},\qquad Z(0)=0.$$

It follows that

$$\frac{\partial\mathcal{S}}{\partial\psi}(0,0,\tau)=\mathcal{S}(\tau)\int_0^T\mathcal{S}^{-1}(s)\Big(D_{xx}F\big(u_0(s)\big)\zeta_1(s)-D_xF\big(u_0(s)\big)\Big)\mathcal{S}(s)ds. \tag{4.14}$$

Let $\Delta =: \big(D_{xx}F(u_0)\zeta_1-D_xF(u_0)\big)$. Then

$$\begin{aligned}\operatorname{trace}\frac{\partial\mathcal{S}}{\partial\psi}(0,0,T)&=\sum_{i=0}^{3}\int_0^T\langle\mathcal{S}(T)\mathcal{S}^{-1}(s)\Delta(s)\mathcal{S}(s)\zeta_i(0)\mid\zeta_i^*(0)\rangle\,ds\\&=\sum_{i=0}^{3}\int_0^T\langle\Delta(s)\mathcal{S}(s)\zeta_i(0)\mid{}^t\mathcal{S}^{-1}(s){}^t\mathcal{S}(T)\zeta_i^*(0)\rangle\,ds.\end{aligned}$$

One can easily check that $\mathcal{S}(s)\mathcal{S}^{-1}(T)=\mathcal{S}(s-T)$, and that ${}^t\mathcal{S}^{-1}(s)$ is the fundamental matrix of the adjoint equation ${}^t\mathcal{A}z=0$. As a consequence,

$$\begin{aligned}\mathcal{S}(s)\zeta_0(0)&=\zeta_0(s),\\\mathcal{S}(s)\zeta_1(0)&=\zeta_1(s)+T\zeta_0(s),\\\mathcal{S}(s)\zeta_2(0)&=\zeta_2(s)+T\zeta_1(s)+T^2\zeta_0(s)/2,\\\mathcal{S}(s)\zeta_3(0)&=\zeta_3(s)+T\zeta_2(s)+T^2\zeta_1(s)/2+T^3\zeta_0(s)/6,\\{}^t\mathcal{S}(s)^{-1}\zeta_0^*(0)&=\zeta_0^*(s)-T\zeta_1^*(s)+T^2\zeta_2^*(s)/2-T^3\zeta_3^*(s)/6,\\{}^t\mathcal{S}(s)^{-1}\zeta_1^*(0)&=\zeta_1^*(s)-T\zeta_2^*(s)+T^2\zeta_3^*(s)/2,\\{}^t\mathcal{S}(s)^{-1}\zeta_2^*(0)&=\zeta_2^*(s)-T\zeta_3^*(s),\\{}^t\mathcal{S}(s)^{-1}\zeta_3^*(0)&=\zeta_3^*(s).\end{aligned}$$

And we obtain

$$\begin{aligned}\operatorname{trace}\frac{\partial\mathcal{S}}{\partial\psi}(0,0,T)=&\,T^4\langle\Delta\zeta_0\mid\zeta_3^*\rangle_T/6+T^3\big(\langle\Delta\zeta_0\mid\zeta_2^*\rangle_T+\langle\Delta\zeta_1\mid\zeta_3^*\rangle_T\big)/2\\&+T^2\big(\langle\Delta\zeta_0\mid\zeta_1^*\rangle_T+\langle\Delta\zeta_1\mid\zeta_2^*\rangle_T+\langle\Delta\zeta_2\mid\zeta_3^*\rangle_T\big)\\&+T\big(\langle\Delta\zeta_0\mid\zeta_0^*\rangle_T+\langle\Delta\zeta_1\mid\zeta_1^*\rangle_T+\langle\Delta\zeta_2\mid\zeta_2^*\rangle_T+\langle\Delta\zeta_3\mid\zeta_3^*\rangle_T\big).\end{aligned}$$

Due to the reversibility, the scalar product in factor of T^3 and T are zero, and it remains that

$$\begin{aligned}\operatorname{trace}\frac{\partial\mathcal{S}}{\partial\psi}(0,0,T)=&\,T^4\langle\Delta\zeta_0\mid\zeta_3^*\rangle_T/6\\&+T^2\big(\langle\Delta\zeta_0\mid\zeta_1^*\rangle_T+\langle\Delta\zeta_1\mid\zeta_2^*\rangle_T+\langle\Delta\zeta_2\mid\zeta_3^*\rangle_T\big).\end{aligned}\tag{4.15}$$

Now, let us compute the coefficient α of the normal form.

Let $x=u_0+\psi\zeta_1+A\zeta_2+B\zeta_3+\Phi(\tau,y)$, then up to the order 2

$$\begin{aligned}F(x)=&\,\zeta_0+\psi D_xF(u_0)\zeta_1+AD_xF(u_0)\zeta_2+BD_xF(u_0)\zeta_3\\&+D_xF(u_0)\Phi+D_{x,x}F(u_0)(y,y)/2\end{aligned}\tag{4.16}$$

$$\begin{aligned}\frac{dx}{d\tau}=&\,\zeta_0+\psi\big(D_xF(u_0)\zeta_1-\zeta_0\big)+AD_xF(u_0)\zeta_2+BD_xF(u_0)\zeta_3\\&+\alpha\psi^2\zeta_2+\alpha A\psi\zeta_3+\frac{\partial\Phi}{\partial\tau}+A\frac{\partial\Phi}{\partial\psi}+B\frac{\partial\Phi}{\partial A}.\end{aligned}\tag{4.17}$$

If we expand

$$n(y) = \sum n_{i,j,k}\psi^i A^j B^k, \qquad \Phi(\tau, y) = \sum \Phi_{i,j,k}(\tau)\psi^i A^j B^k,$$

the identification of the terms of order ψ^2 in the equation

$$\big(1+\psi+n(y)\big)\frac{dx}{d\tau} = F(x), \tag{4.18}$$

leads to

$$\begin{aligned} -\Phi'_{2,0,0} + D_xF(u_0)\Phi_{2,0,0} =& n_{2,0,0}\zeta_0 - \zeta_0 + \alpha\zeta_2 + D_xF(u_0)\zeta_1 \\ & - D_{x,x}F(u_0)(\zeta_1,\zeta_1)/2. \end{aligned} \tag{4.19}$$

The solvability condition of this equation reads

$$\langle D_xF(u_0)\zeta_1 - D_{x,x}F(u_0)(\zeta_1,\zeta_1)/2 \mid \zeta_3^*\rangle_T = 0,$$

and is satisfied due to the reversibility. In the same way, the identification of the terms of order $A\psi$ leads to

$$\begin{aligned} -\Phi'_{1,1,0} + D_xF(u_0)\Phi_{1,1,0} =& n_{1,1,0}\zeta_0 + \alpha\zeta_3 + D_xF(u_0)\zeta_2 \\ & - D_{x,x}F(u_0)(\zeta_1,\zeta_2) + 2\Phi_{2,0,0}. \end{aligned} \tag{4.20}$$

The solvability condition now reads

$$\alpha + \langle D_xF(u_0)\zeta_2 - D_{x,x}F(u_0)(\zeta_1,\zeta_2) \mid \zeta_3^*\rangle_T + 2\,\langle \Phi_{2,0,0} \mid \zeta_3^*\rangle_T = 0.$$

From (4.8) and (4.19), we get

$$\langle \Phi_{2,0,0} \mid \zeta_3^*\rangle_T = \alpha + \langle D_xF(u_0)\zeta_1 - D_{x,x}F(u_0)(\zeta_1,\zeta_1)/2 \mid \zeta_2^*\rangle_T,$$

and compute

$$\begin{aligned} 3\alpha =& \langle D_{x,x}F(u_0)(\zeta_1,\zeta_1) - 2D_xF(u_0)\zeta_1 \mid \zeta_2^*\rangle_T \\ & + \langle D_{x,x}F(u_0)(\zeta_1,\zeta_2) - D_xF(u_0)\zeta_2 \mid \zeta_3^*\rangle_T. \end{aligned} \tag{4.21}$$

It follows that

$$\begin{aligned} \operatorname{trace}\frac{\partial\mathcal{S}}{\partial\psi}(0,0,T) - 3\alpha T^2 =& T^4\big(\langle D_{x,x}F(u_0)(\zeta_1,\zeta_0) \mid \zeta_3^*\rangle_T - \langle D_xF(u_0)\zeta_0 \mid \zeta_3^*\rangle_T\big)/6 \\ & + T^2\big(\langle D_{x,x}F(u_0)(\zeta_1,\zeta_0) \mid \zeta_1^*\rangle_T - \langle D_xF(u_0)\zeta_0 \mid \zeta_1^*\rangle_T \\ & \quad + \langle D_xF(u_0)\zeta_1 \mid \zeta_2^*\rangle_T\big). \end{aligned}$$

Moreover, writing that $\zeta_2^* = {}^t\mathcal{A}\zeta_1^*$, one gets

$$\begin{aligned} \langle D_xF(u_0)\zeta_1 \mid \zeta_2^*\rangle_T =& \left\langle D_xF(u_0)\zeta_1 \;\middle|\; \frac{d\zeta_1^*}{d\tau} + {}^tD_xF(u_0)\zeta_1^* \right\rangle_T \\ =& -\langle D_{x,x}F(u_0)(\zeta_1,\zeta_0) \mid \zeta_1^*\rangle_T - \left\langle D_xF(u_0)\frac{d\zeta_1}{d\tau} \;\middle|\; \zeta_1^* \right\rangle_T \\ & + \langle D_xF(u_0)\zeta_1 \mid {}^tD_xF(u_0)\zeta_1^*\rangle_T \\ =& -\langle D_{x,x}F(u_0)(\zeta_1,\zeta_0) \mid \zeta_1^*\rangle_T \\ & + \left\langle -\frac{d\zeta_1}{d\tau} + D_xF(u_0)\zeta_1 \;\middle|\; \zeta_1^* \right\rangle_T \\ =& -\langle D_{x,x}F(u_0)(\zeta_1,\zeta_0) \mid \zeta_1^*\rangle_T \\ & + \langle D_xF(u_0)\zeta_0 \mid \zeta_1^*\rangle_T, \end{aligned}$$

i.e.,

$$\langle D_{x,x}F(u_0)(\zeta_1,\zeta_0) \mid \zeta_1^*\rangle_T - \langle D_xF(u_0)\zeta_0 \mid \zeta_1^*\rangle_T + \langle D_xF(u_0)\zeta_1 \mid \zeta_2^*\rangle_T = 0.$$

Writing that $\zeta_3^* = {}^t\mathcal{A}\zeta_2^*$, we obtain similarly

$$\langle D_{x,x}F(u_0)(\zeta_1,\zeta_0) \mid \zeta_3^*\rangle_T = \langle D_xF(u_0)\zeta_0 \mid \zeta_3^*\rangle_T\,.$$

As a consequence,

$$\text{trace}\,\frac{\partial\mathcal{S}}{\partial\psi}(0,0,T) = 3\alpha T^2. \tag{4.22}$$

We finally have stated the following.

THEOREM. *Generically if* $[\alpha \neq 0]$, *the solution* u_0 *and, therefore, the families of periodic and homoclinic orbits of* (2.1), *are stable under small reversible perturbations.*

5. Application to the reversible 1:1 resonance. Let us consider a four-dimensional differential equation

$$\frac{dX}{dt} = \mathcal{F}(\mu, X), \tag{5.1}$$

with a fixed point at the origin such that the linearized operator for $\mu = 0$ has a 1:1 resonance, i.e., such that

$$D_X\mathcal{F}(0,0) = \begin{pmatrix} 0 & -\omega_0 & 1 & 0 \\ \omega_0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\omega_0 \\ 0 & 0 & \omega_0 & 0 \end{pmatrix}. \tag{5.2}$$

In [5], the authors showed that if in addition the system is reversible, (5.1) can be put into *a normal form* that writes

$$\frac{dX}{dt} = \mathcal{F}'(\mu, X), \tag{5.3}$$

with

$$X = \begin{pmatrix} A \\ B \end{pmatrix}, \qquad A, B \in \mathbb{C},$$

$$\mathcal{F}'(\mu,X) = \begin{pmatrix} i\omega_0 A + B + iA\mathcal{P}\left(\mu, |A|^2, \frac{i}{2}(A\bar{B} - \bar{A}B)\right) \\ i\omega_0 B + iB\mathcal{P}\left(\mu, |A|^2, \frac{i}{2}(A\bar{B} - \bar{A}B)\right) + A\mathcal{Q}\left(\mu, |A|^2, \frac{i}{2}(A\bar{B} - \bar{A}B)\right) \end{pmatrix},$$

and $\mathcal{P}$, $\mathcal{Q}$ polynomial in their arguments such that $\mathcal{P}(0,0,0) = \mathcal{Q}(0,0,0) = 0$.

The system (5.3) is rotationally invariant, i.e.,

$$\mathcal{F}'(\mu, R_\phi X) = R_\phi \mathcal{F}'(\mu, X), \quad \text{where } R_\phi = e^{\phi L}, \quad \text{and} \quad L\begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} iA \\ iB \end{pmatrix},$$

and admits two first integrals:

$$\begin{cases} \mathcal{K} = \frac{i}{2}(A\bar{B} - \bar{A}B), \\ \mathcal{H} = |B|^2 - \mathcal{G}(\mu, |A|^2, \mathcal{K}), \end{cases} \quad \text{where } \mathcal{G}(\mu, u, \mathcal{K}) =: \int_0^u \mathcal{Q}(\mu, s, \mathcal{K})ds. \tag{5.4}$$

If we define

$$x =: \Re e A e^{-i\omega_0 t} \quad y =: \Im m A e^{-i\omega_0 t}, \qquad \bar{X} = \begin{pmatrix} x \\ y \\ z \\ \mathcal{T} \end{pmatrix},$$
$$z =: \Re e B e^{-i\omega_0 t} \quad \mathcal{T} =: \Im m B e^{-i\omega_0 t},$$

then X is a solution of (5.3) if and only if $\bar{X}$ is a solution of

$$\frac{d\bar{X}}{dt} = \bar{\mathcal{F}}(\mu, \bar{X}), \tag{5.5}$$

with

$$\bar{\mathcal{F}}(\mu, \bar{X}) = \mathcal{F}'(\mu, \bar{X}) - \omega_0 L\bar{X} = \begin{pmatrix} - & y\mathcal{P}(\mu, u, \kappa) + z \\ & x\mathcal{P}(\mu, u, \kappa) + \mathcal{T} \\ - & \mathcal{T}\mathcal{P}(\mu, u, \kappa) + x\mathcal{Q}(\mu, u, \kappa) \\ & z\mathcal{P}(\mu, u, \kappa) + y\mathcal{Q}(\mu, u, \kappa) \end{pmatrix},$$

and $u = x^2 + y^2$, $\kappa = x\mathcal{T} - yz$.

It is proved in [5] that the system (5.5) admits periodic solutions of type

$$\bar{X}_\kappa(t) = R_{\alpha_\kappa t}\bar{X}_\kappa, \tag{5.6}$$

where

$$\bar{X}_\kappa = \begin{pmatrix} r_0 \\ 0 \\ 0 \\ r_1 \end{pmatrix}$$

satisfies

$$\mathcal{F}'(\mu, \bar{X}_\kappa) = (\omega_0 + \alpha_\kappa)L\bar{X}_\kappa. \tag{5.7}$$

The linearized equation around the periodic orbit $\mathcal{O} =: \{\bar{X}_\kappa(t)/t \in \mathbb{R}\}$ is

$$\frac{dZ}{dt} - \mathcal{L}_\kappa(t)Z = 0, \tag{5.8}$$

with $\mathcal{L}_\kappa(t) = D_{\bar{X}}\bar{\mathcal{F}}\big(\mu, \bar{X}_\kappa(t)\big) = D_X\mathcal{F}'\big(\mu, \bar{X}_\kappa(t)\big) - \omega_0 L$. If we let Z_λ be an eigenvector of $\mathcal{L}_\kappa =: D_X\mathcal{F}'(\mu, \bar{X}_\kappa) - (\omega_0 + \alpha_\kappa)L$, belonging to the eigenvalue λ, then $Z_\lambda(t) =: R_{\alpha_\kappa t}Z_\lambda$ is a solution of

$$\frac{dZ}{dt} - \mathcal{L}_\kappa(t)Z = \lambda Z.$$

This proves that the Floquet exponents of (5.8) are mod. $\iota\alpha_\kappa$, the eigenvalues of $\mathcal{L}_\kappa$. Moreover, one can compute

$$\mathcal{L}_\kappa = \begin{pmatrix} 0 & \frac{r_1}{r_0} & 1 & 0 \\ a + c & 0 & 0 & \frac{r_0}{r_1}c \\ \lambda^2 - \frac{r_1}{r_0}(a - b + c) & 0 & 0 & b - c \\ 0 & -\left(\frac{r_1}{r_0}\right)^2 & -\frac{r_1}{r_0} & 0 \end{pmatrix}, \tag{5.9}$$

where

$$a = 2\left(\mathcal{P}'_u r_0^2 - \frac{r_1}{r_0}\right), \qquad b = \mathcal{Q}'_\kappa r_0^2 + 2\frac{r_1}{r_0},$$

$$c = \mathcal{P}'_\kappa r_0 r_1 + \frac{r_1}{r_0}, \qquad \lambda^2 = \mathcal{Q}'_u r_0^2 - 2\left(\frac{r_1}{r_0}\right)^2.$$

When $\lambda \neq 0$, the eigenvectors of $\mathcal{L}_{\mathcal{K}}$ are

$$X_{\pm\lambda} = \begin{pmatrix} \lambda \\ \pm a \\ \pm\lambda^2 \mp \frac{r_1}{r_0} a \\ \frac{r_1}{r_0}\lambda \end{pmatrix} \begin{matrix} \text{belonging to} \\ \text{the eigenvalue } \pm\lambda \end{matrix}, \quad LX_{\mathcal{K}} = \begin{pmatrix} 0 \\ r_{0_{\mathcal{K}}} \\ -r_{1_{\mathcal{K}}} \\ 0 \end{pmatrix} \begin{matrix} \text{belonging to} \\ \text{the eigenvalue zero.} \end{matrix}$$

Notice that the eigenvenvalue zero is not semisimple because, differentiating (5.7) with respect to $\mathcal{K}$, one obtains

$$\mathcal{L}_{\mathcal{K}} \frac{dX_{\mathcal{K}}}{d\mathcal{K}} = \frac{d\alpha_{\mathcal{K}}}{d\mathcal{K}} LX_{\mathcal{K}}. \tag{5.10}$$

According to (16) and (17) of [5] we have

$$u_{\mathcal{K}}^2 \mathcal{Q}(\mu, u_{\mathcal{K}}, \mathcal{K}) + \mathcal{K}^2 = 0, \qquad \alpha_{\mathcal{K}} = \mathcal{P}(\mu, u_{\mathcal{K}}, \mathcal{K}) + \frac{\mathcal{K}}{u_{\mathcal{K}}}. \tag{5.11}$$

We can deduce that

$$\frac{du_{\mathcal{K}}}{d\mathcal{K}} = \frac{-2\mathcal{K}}{\lambda^2 u_{\mathcal{K}}}, \qquad \frac{d\alpha_{\mathcal{K}}}{d\mathcal{K}} = \frac{1}{u_{\mathcal{K}}}\left(1 - \frac{a\mathcal{K}}{\lambda^2 u_{\mathcal{K}}}\right). \tag{5.12}$$

It is proved in [5] that in the supercritical case $[(\partial\mathcal{Q}/\partial u)(0,0,0) > 0]$, when $\mu > 0$, there exists a point (the Eckhaus point E of Fig. 3 in [5]), where $\lambda = 0$. As we get to this point, $\lambda \to 0$, $d\alpha_{\mathcal{K}}/d\mathcal{K} \to -\infty$ and all eigenvectors $X_{\pm\lambda}$ tend towards $LX_E =: LX_{\mathcal{K}_E}$. Moreover, when $\lambda = 0$ $[\mathcal{K} = \mathcal{K}_E]$, one can check that all eigenvalues of $\mathcal{L}_E =: \mathcal{L}_{\mathcal{K}_E}$ vanish and that, as $ab \neq 0$, $\mathrm{Ker}(\mathcal{L}_E)$ is one-dimensional, i.e., *the Jordan block belonging to the eigenvalue zero is four-dimensional.*

We are now going to define a rescaling so that all these "critical" periodic orbits we found for $\mu > 0$ appear for $\nu = 0$. Using the same notation, as in [5], we expand

$$\begin{aligned} \mathcal{P}(\mu, u, \mathcal{K}) &= \quad p_1\mu + p_2 u + p_3\mathcal{K} + \cdots \\ \mathcal{Q}(\mu, u, \mathcal{K}) &= -q_1\mu + q_2 u + q_3\mathcal{K} + \cdots \quad [q_1, q_2 > 0]. \end{aligned}$$

After the change of parameter $\nu^2 =: q_1\mu/3$, we get at the first order,

$$\begin{gathered} \mathcal{K}_E = \frac{2}{q_2}\left(\frac{q_1\mu}{3}\right)^{3/2} + \cdots = \frac{2}{q_2}\nu^3 + \cdots, \\ u_E = \frac{2}{q_2}^{1/3} \mathcal{K}_E^{2/3} + \cdots = \frac{2}{q_2}\nu^2 + \cdots, \\ a = -2\nu + \cdots, \ b = 2\nu + \cdots, \ c = \nu + \cdots, \end{gathered}$$

and

$$\mathcal{L}_E = \begin{pmatrix} 0 & \nu & 1 & 0 \\ -\nu & 0 & 0 & 1 \\ 3\nu^2 & 0 & 0 & \nu \\ 0 & -\nu^2 & -\nu & 0 \end{pmatrix}. \tag{5.13}$$

If we make the rescaling,

(5.14)

$$\hat{X} = \begin{pmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \\ \hat{\tau} \end{pmatrix} = \begin{pmatrix} \frac{1}{\nu}x \\ \frac{1}{\nu}y \\ \frac{1}{\nu^2}z \\ \frac{1}{\nu^2}\tau \end{pmatrix}, \quad \hat{t} =: \nu t, \quad \hat{\mathcal{F}}(\nu, \hat{X}) = \begin{pmatrix} \frac{1}{\nu^2}\left(-y\mathcal{P}(\mu, u, \mathcal{K}) + z\right) \\ \frac{1}{\nu^2}\left(x\mathcal{P}(\mu, u, \mathcal{K}) + \tau\right) \\ \frac{1}{\nu^3}\left(-\tau\mathcal{P}(\mu, u, \mathcal{K}) + x\mathcal{Q}(\mu, u, \mathcal{K})\right) \\ \frac{1}{\nu^3}\left(z\mathcal{P}(\mu, u, \mathcal{K}) + y\mathcal{Q}(\mu, u, \mathcal{K})\right) \end{pmatrix},$$

then $\bar{X}$ is a solution of (5.5) if and only if $\hat{X}$ is a solution of

$$\frac{d\hat{X}}{d\hat{t}} = \hat{\mathcal{F}}(\nu, \hat{X}). \tag{5.15}$$

Note that

$$X_E = \begin{pmatrix} r_{0_E} \\ 0 \\ 0 \\ r_{1_E} \end{pmatrix}, \qquad \hat{X}_E = \sqrt{\frac{2}{\mathcal{Q}_2}} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

$$\mathcal{Q}(\mu, \mathcal{U}, \kappa) = \nu^2(\mathcal{Q}_2\hat{\mathcal{U}} - 3) + O(\nu^3), \quad \mathcal{P}(\mu, \mathcal{U}, \kappa) = O(\nu^2), \quad \alpha_\kappa = \nu + O(\nu^2), \tag{5.16}$$

$$\hat{\mathcal{F}}(0, \hat{X}) = \begin{pmatrix} \hat{\mathcal{Z}} \\ \hat{\mathcal{T}} \\ \hat{\mathcal{X}}(\mathcal{Q}_2\hat{\mathcal{U}} - 3) \\ \hat{\mathcal{Y}}(\mathcal{Q}_2\hat{\mathcal{U}} - 3) \end{pmatrix}.$$

So now the solutions $\bar{X}_E(t)$ of (5.5) correspond to the solution $\hat{X}_E(\hat{t}) = R_{\hat{t}}\hat{X}_E$ of (5.13), when $\nu = 0$. Moreover, our rescaling has been chosen in such a way that

$$\hat{\mathcal{L}}_E = \begin{pmatrix} 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 3 & 0 & 0 & 1 \\ 0 & -1 & -1 & 0 \end{pmatrix}. \tag{5.17}$$

The matrix $\hat{\mathcal{L}}_E$ is nilpotent of order 4, and a good choice of a Jordan basis for $\hat{\mathcal{L}}_E$ is as follows:

$$\hat{X}_0 = \frac{1}{\sqrt{2\mathcal{Q}_2}} \begin{pmatrix} 0 \\ 2 \\ -2 \\ 0 \end{pmatrix}, \qquad \hat{X}_1 = \frac{1}{\sqrt{2\mathcal{Q}_2}} \begin{pmatrix} -1 \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

$$\hat{X}_2 = \frac{1}{\sqrt{2\mathcal{Q}_2}} \begin{pmatrix} 0 \\ 0 \\ -1 \\ 0 \end{pmatrix}, \qquad \hat{X}_3 = \frac{1}{\sqrt{2\mathcal{Q}_2}} \begin{pmatrix} -\frac{1}{4} \\ 0 \\ 0 \\ -\frac{1}{4} \end{pmatrix}.$$

If we let $\zeta_i(\hat{t}) = R_{\hat{t}}\hat{X}_i$, the relations $\hat{\mathcal{L}}_E\hat{X}_i = \hat{X}_{i-1}$ lead to

$$\begin{gathered} \zeta_0(\hat{t}) = \frac{dR_{\hat{t}}\hat{X}_E}{d\hat{t}}(\hat{t}), \\ \left(-\frac{d}{d\hat{t}} + \hat{\mathcal{L}}_E(\hat{t})\right)\zeta_i(\hat{t}) = \zeta_{i-1}(\hat{t}), \qquad i = 1, 2, 3. \end{gathered} \tag{5.18}$$

The adjoint basis of $(\hat{X}_0, \hat{X}_1, \hat{X}_2, \hat{X}_3)$ is

$$\hat{X}_0^* = \sqrt{2\mathcal{Q}_2} \begin{pmatrix} 0 \\ \frac{1}{2} \\ 0 \\ 0 \end{pmatrix}, \qquad \hat{X}_1^* = \sqrt{2\mathcal{Q}_2} \begin{pmatrix} -\frac{1}{2} \\ 0 \\ 0 \\ \frac{1}{2} \end{pmatrix},$$

$$\hat{X}_2^* = \sqrt{2\mathcal{Q}_2} \begin{pmatrix} 0 \\ -1 \\ -1 \\ 0 \end{pmatrix}, \qquad \hat{X}_3^* = \sqrt{2\mathcal{Q}_2} \begin{pmatrix} -2 \\ 0 \\ 0 \\ -2 \end{pmatrix}.$$

If we let $\zeta_i^*(\hat{t}) = R_{-\hat{t}}\hat{X}_i^*$, one can check that

$$\left(\frac{d}{d\hat{t}} + {}^t\hat{\mathcal{L}}_E(\hat{t})\right)\zeta_i^*(\hat{t}) = \zeta_{i+1}^*(\hat{t}), \qquad i = 0 \cdots 3. \tag{5.19}$$

Due to the expression of the ζ_i and ζ_i^* and the rotational invariance of $\hat{F}$, (4.21) takes the "simpler" form

$$\begin{aligned} 3\alpha = &\left\langle D_{\hat{x},\hat{x}}\hat{F}(\hat{X}_E)(\hat{X}_1,\hat{X}_1) - 2D_{\hat{x}}\hat{F}(\hat{X}_E)\hat{X}_1 \mid \hat{X}_2^*\right\rangle, \\ &\left\langle D_{\hat{x},\hat{x}}\hat{F}(\hat{X}_E)(\hat{X}_1,\hat{X}_2) - D_{\hat{x}}\hat{F}(\hat{X}_E)\hat{X}_2 \mid \hat{X}_3^*\right\rangle, \end{aligned}$$

with, according to (5.16),

$$D_{\hat{x}}\hat{F}(\hat{X}_E)\bar{X} = \begin{pmatrix} \hat{z} \\ \hat{T} \\ 3\hat{x} \\ -\hat{y} \end{pmatrix}, \qquad D_{\hat{x},\hat{x}}\hat{F}(\hat{X}_E)(\hat{X}_1,\hat{X}) = \begin{pmatrix} 0 \\ 0 \\ -3\hat{x} \\ -\hat{y} \end{pmatrix}.$$

This allows us to compute

$$\alpha = -4. \tag{5.20}$$

So, the study of the reversible 1:1 resonance enters into the frame of §§3 and 4 with $\alpha = -4$: in our problem the origin in the (H, K) plane is just the same as the point E in [5].

Acknowledgments. The author would like to thank André Vanderbauwhede for the attention he has paid to this work, and the referee for suggesting a great improvement to an earlier result.

REFERENCES

[1] V. I. Arnold, *Geometrical methods in the theory of ordinary differential equations*, Grundlehren Math. Wiss. 250, Springer-Verlag, New York, Heidelberg, Berlin, 1983.

[2] C. Elphick, E. Tirapegui, M. Brachet, P. Coullet, and G. Iooss, *A simple global characterization for normal forms of singular vector fields*, Physica, 29 D (1987), pp. 95–127.

[3] J. Hale, *Ordinary Differential Equations*, John Wiley, New York, 1969.

[4] G. Iooss, *Global characterization of the normal form for a vector field near a closed orbit*, J. Differential Equations, 76 (1988), pp. 47–76.

[5] G. Iooss and M. C. Pérouème, *Perturbed homoclinic solutions in reversible* 1:1 *resonance vector field*, J. Differential Equations, 102 (1993), pp. 62–88.

[6] A. Vanderbauwhede, *Branching of periodic solutions in time-reversible systems*, in Geometry and Analysis in Non Linear Dynamics, H. Brœr and F. Takens, eds., Longman, London, 1991, to appear.

SIAM J. MATH. ANAL.
Vol. 24, No. 6, pp. 1597–1621, November 1993

010

CONTINUED FRACTION REPRESENTATIONS OF MAXIMAL AND MINIMAL SOLUTIONS OF A DISCRETE MATRIX RICCATI EQUATION*

CALVIN AHLBRANDT[†]

Abstract. Explicit constructions are given for the "maximal" solution W_n^+ and the "minimal" solution W_n^- of the discrete matrix Riccati equation

$$W_{n+1} = A_n + W_n(W_n + C_{n-1})^{-1}C_{n-1}$$

given "disconjugacy" of the associated self-adjoint linear difference equation

$$-\Delta(C_{n-1}\Delta x_{n-1}) + A_n x_n = 0.$$

These constructions provide characterizations of the recessive solutions at $-\infty$ and ∞ of the linear equation. In the special case where A_n and C_n are positive definite, the solutions W_n^+ and W_n^- have the simple continued fraction representations (using $D_n \equiv C_n^{-1}$)

$$W_n^+ = \lim_{M\to-\infty} \quad A_{n-1} + \frac{I}{D_{n-2}} + \frac{I}{A_{n-2}} + \frac{I}{D_{n-3}} + \cdots + \frac{I}{A_{M+1}} + \frac{I}{D_M}$$

and

$$-\frac{I}{W_n^-} = \lim_{N\to\infty} \quad D_{n-1} + \frac{I}{A_n} + \frac{I}{D_n} + \frac{I}{A_{n+1}} + \cdots + \frac{I}{A_{N-1}} + \frac{I}{D_{N-1}},$$

respectively. The periodic coefficient case provides a matrix extension of a result of Galois. Matrix extensions are also provided for continued fraction results of Euler and Pincherle.

Key words. matrix continued fractions, discrete Riccati equations, maximal solution, difference equations

AMS subject classifications. 39A10, 30B70, 11Y65

1. Introduction.

As in Ahlbrandt and Hooker, [5] and [6], consider a vector difference equation

$$-\Delta(C_{n-1}\Delta x_{n-1}) + A_n x_n = 0, \tag{1.1}$$

where A_n and C_n are $r \times r$ real symmetric matrices with C_n nonsingular for all integers n. The associated matrix difference equation is

$$-\Delta(C_{n-1}\Delta X_{n-1}) + A_n X_n = 0. \tag{1.2}$$

By expanding the forward differences, $(\Delta u_k = u_{k+1} - u_k)$, these equations have the form of symmetric, three term recurrence relations

$$-C_n x_{n+1} - C_{n-1}x_{n-1} + B_n x_n = 0 \tag{1.3}$$

*Received by the editors January 31, 1990; accepted for publication (in revised form) September 21, 1992.

[†]Department of Mathematics, University of Missouri, Columbia, Missouri 65211.

and

$$-C_nX_{n+1} - C_{n-1}X_{n-1} + B_nX_n = 0 \tag{1.4}$$

with $B_n = C_n + C_{n-1} + A_n$.

The corresponding discrete Riccati equation is

$$W_{n+1} = A_n + W_n(W_n + C_{n-1})^{-1}C_{n-1}. \tag{1.5}$$

Solutions X_n of (1.2) and W_n of (1.5) are related by

$$W_n = (C_{n-1}\Delta X_{n-1})X_{n-1}^{-1}. \tag{1.6}$$

In previous joint work with Hooker [6] it was shown that a hypothesis of "eventual disconjugacy"of (1.1) implies existence of a recessive solution of (1.2) at ∞ and a corresponding "eventually minimal" Hermitian solution, W_n^-, of the Riccati equation (1.5). An analogous treatment will now be given for the recessive solution at $-\infty$ and the corresponding "primordial maximal" solution, W_n^+, of (1.5). This language is chosen because solutions which start outside the "interval" $[W_n^-, W_n^+]$ can be extensible. This contrasts with the continuous case where no solution can extend to the right if it starts below $W^-(t)$ nor extend to the left if it starts above $W^+(t)$. However, as a consequence of results of [6] and Theorem 6.1, below, it follows that under the disconjugacy hypothesis, any Hermitian solution W_n of (1.5) on $(-\infty, \infty)$ can violate the condition $W_n^- \leq W_n \leq W_n^+$ a finite number of times at most. The Riccati interpretation of the Reid constructions [36] of the recessive solutions will prove convergence of generalized, continued fraction representations for W_n^+ and W_n^-. These results for W_n^- also improve the previous results given in [6]. The general theory is then applied to special cases of periodic coefficients, positive-definite coefficients, and constant positive-definite coefficients. The latter case adds perspective to the results of [6], where the constant coefficient results were obtained by time-reversal methods. The results in the periodic case are related to a study of periodic and reverse periodic continued fractions by Merkes and Scott [28]. In particular, the result of Galois [18] cited there [28, Cor. 1, p. 26] suggested the aesthetic notational relationship of W_n^+ to W_n^- given in the abstract. Furthermore, it shows the dependence of the "minimal" solution upon present and future coefficients, whereas the "maximal" solution depends upon all past coefficients. Ryde [38, Chap. V, pp. 67–82] discussed relationships between finite continued fractions and their reciprocals (he called them "inverse" continued fractions).

Arscott [7], [8] also used the type of construction used by Reid [36] and Gautschi [19] in the study of the recessive solution of three term recursions arising in eigenvalue problems for periodic solutions of Mathieu's differential equation. Noncommutative continued fractions have been studied by Fair [15]–[17], and by Denk and Riederle [13]. Reference [13] includes an extensive bibliography. The question of "oscillation" of (1.2) was recently investigated by Chen and Erbe [10] and by Peterson and Ridenhour [32]. Peterson and Ridenhour [32, eq. (8)] point out that the Riccati equation can also be written in the form

$$\Delta W_n = A_n - W_n(W_n + C_{n-1})^{-1}W_n. \tag{1.7}$$

This shows that the iteration (1.5) carries Hermitian W_n to Hermitian W_{n+1}.

This work was motivated by more complicated Riccati equations arising in the discrete regulator problem; see, for example, Vaughan [44].

Theorem 8.1 makes possible an "analytic theory" formulation of matrix continued fractions. A theory of "symplectic continued fractions" is introduced in §13. That theory allows extensions of theorems of Euler and Pincherle to the matrix case. Theorem 13.1 equates convergence of a matrix continued fraction with convergence of an associated series. Theorem 13.2 establishes equivalence of existence of a recessive solution at ∞ with convergence of a continued fraction. It is to be noted that §13 requires no sign conditions on the coefficients, other than C_n nonsingular, and assumes no Sturmian theory. The experienced reader should start by reading §13.

2. Terminology and basic propositions.

The notation U' denotes the conjugate transpose of U.

PROPOSITION 2.1. *If U_n and V_n are solutions of* (1.2), *then the bracket function*

$$\begin{aligned}\{U_n, V_n\} &= U_n' C_{n-1}\Delta V_{n-1} - [C_{n-1}\Delta U_{n-1}]'V_n \\ &= U_{n-1}' C_{n-1} V_n - U_n' C_{n-1} V_{n-1}\end{aligned} \tag{2.1}$$

has $\Delta\{U_n, V_n\} = 0$; i.e., $\{U_n, V_n\}$ is constant. Also,

$$\{U_n, V_n\}' = -\{V_n, U_n\}. \tag{2.2}$$

Proposition 2.1 also holds for vector solutions. Vector solutions u and v are called *conjoined* [35] if $\{u_n, v_n\} = 0$. A vector solution u is called *self-conjoined* if $\{u_n, u_n\} = 0$. A matrix solution X is called *prepared* [23] (or self-conjoined) if $\{X_n, X_n\} \equiv 0$.

PROPOSITION 2.2 (variation of parameters). *If X_n is a solution of a homogeneous equation $\Delta X_{n-1} + F_n X_n = 0$ with X_n nonsingular for $M < n \leq N$, then Y_n is a solution of the nonhomogeneous equation $\Delta Y_{n-1} + F_n Y_n = G_n$ if and only if Y_n has the form $Y_n = X_n V_n$, $M < n \leq N$ with V_n a solution of the first-order equation*

$$\Delta V_{n-1} = X_{n-1}^{-1} G_n, \qquad M < n - 1 \leq N. \tag{2.3}$$

PROPOSITION 2.3 (reduction of order). *If X_n is any prepared solution of* (1.2) *with X_n nonsingular for $M < n \leq N$, then Y_n is a solution of* (1.2) *if and only if it has the form*

$$Y_n = X_n[P - S_{n,N}(X)Q], \qquad M < n \leq N \tag{2.4}$$

where

$$P = X_N^{-1} Y_N, \qquad Q = \{X_n, Y_n\} \tag{2.5}$$

and

$$S_{n,N}(X) = \begin{cases} 0, & n = N, \\ \sum_{k=n+1}^{N} (X_k' C_{k-1} X_{k-1})^{-1}, & n < N. \end{cases} \tag{2.6}$$

Furthermore, Y_n is prepared if and only if

$$P'Q = Q'P. \tag{2.7}$$

PROPOSITION 2.4. *If U_n is any solution of (1.2), then each of the $2r \times r$ matrices*

$$\mathcal{U}_n = \begin{bmatrix} U_n \\ C_{n-1}\Delta U_{n-1} \end{bmatrix}$$

has the same rank. In particular, if $\mathcal{U}_n$ has full column rank r at one point n, then $\mathcal{U}_n$ has rank r for all n.

A prepared solution U_n such that $\mathcal{U}_n$ has full rank is called a *prepared basis.*

3. Sturmian theory. Jerry Ridenhour and Allan Peterson have pointed out to the author that the definition of conjugate intervals used in [3]–[6] is not appropriate for Theorem 3.1 of [5]. The definition should be as follows: For integers p and q with $p < q$, the intervals $[p, p+1]$ and $[q, q+1]$ are *conjugate* if there exists a self-conjoined vector solution x_n of (1.1) such that

$$x_p' C_p x_{p+1} \le 0 \quad \text{and} \quad x_q' C_q x_{q+1} \le 0 \tag{3.1}$$

with $x_{p+1} \ne 0$ and $x_q \ne 0$. Equation (1.1) is called *disconjugate* on $[M-1, N]$ if the interval $[M-1, N]$ contains no pair of conjugate intervals.

THEOREM 3.1 (see [5]). *Assume A_n and C_n are real symmetric and C_n is nonsingular for all n. Let M and N be integers with $M < N-1$. Then the following conditions are equivalent.*

(i) *If u is a real vector solution of (1.1) on $[M, N]$ with $u_{M-1}' C_{M-1} u_M \le 0$ and $u_M \ne 0$, then*

$$u_n' C_n u_{n+1} > 0 \quad \textit{for } n = M, \ldots, N-1.$$

(ii) *If v is a real vector solution of (1.1) on $[M, N]$ with $v_{N-1}' C_{N-1} v_N \le 0$ and $v_{N-1} \ne 0$, then*

$$v_n' C_n v_{n+1} > 0 \quad \textit{for } n = M-1, \ldots, N-2.$$

(iii) *Equation (1.1) is disconjugate on $[M-1, N]$.*

(iv) *There exists a prepared matrix solution X_n of (1.2) on $[M, N]$ with real entries such that*

$$X_{n-1}' C_{n-1} X_n > 0 \quad \textit{for } n = M, \ldots, N.$$

(v) *There exists a sequence W_n of real symmetric matrices, defined for $n = M, \ldots, N+1$, with $W_n + C_{n-1} > 0$ for $n = M, \ldots, N$ satisfying the matrix Riccati equation (1.5) for $n = M, \ldots, N$.*

(vi) *The quadratic form $\mathcal{J}_2$ defined by*

$$\mathcal{J}_2[\eta] = \sum_M^N (\Delta\eta_{n-1})' C_{n-1} \Delta\eta_{n-1} + \eta_n' A_n \eta_n$$

is positive definite on the class of real vector sequences η with $\eta_{M-1} = 0 = \eta_N$.

Since the quadratic form $\mathcal{J}_2$ has a matrix representation, we have the following.

COROLLARY 3.1. *Disconjugacy on $(-\infty, \infty)$ is equivalent to the condition that the block tridiagonal matrix*

$$\begin{bmatrix} B_M & -C_M & \\ -C_M & \ddots & \ddots \\ & -C_{N-2} & B_{N-1} \end{bmatrix}$$

is positive definite for every M, N, with $M < N - 1$.

Hence this work is conceptually related to Wall's positive-definite continued fractions [45, Chap. IV]. Indeed, this condition will make all inverses exist in our matrix continued fractions. (The referee has pointed out that there is a "related but distinct" paper of MacNerney [27]. That paper extends many of the results of Wall's text [45, Chap. IV] on positive-definite continued fractions to continued fractions of operators.)

THEOREM 3.2 (Sturm separation theorem [6]). *Assume A_n and C_n are real symmetric matrices of order r with C_n nonsingular. If (1.1) is disconjugate on $[M-1, N]$ and U_n is any prepared basis, then the following applies:*

(i) *There exist at most r points m in $[M, N]$ such that U_m is singular; and*

(ii) *There exist at most r points m in $[M, N]$ for which there exist unit vectors $\pi(m)$ such that*

$$(\pi(m))' U_m' C_m U_{m+1} \pi(m) < 0.$$

COROLLARY 3.2. *Suppose (1.1) is disconjugate on $(-\infty, N]$ and U_n is a prepared basis. Then there exists an integer P in $(-\infty, N)$ such that*

$$U_n' C_n U_{n+1} > 0 \quad \text{on } (-\infty, P]. \tag{3.2}$$

In particular, U_n is nonsingular in some neighborhood of $-\infty$.

4. A Hartman-type construction of the recessive solution at $-\infty$.

We assume that (1.1) is disconjugate on some neighborhood of $-\infty$, say $(-\infty, N+1]$. Let X_n be the solution of (1.2) with

$$X_N = I, \qquad X_{N+1} = 0. \tag{4.1}$$

Then $X_n' C_n X_{n+1} > 0$, for $n \le N - 1$, by condition (ii) of Theorem 3.1. Apply Proposition 2.3 with $P = I$ and $Q = -I$ in order to define Y_n by (2.4). Then Y_n is prepared and

$$Y_n = X_n[I + S_{n,N}(X)] \quad \text{for } \le N, \tag{4.2}$$

where $S_{N,N}(X) = 0$ and $S_{n,N}(X) > 0$ for $n < N$. Furthermore, $S_{n,N}(X)$ increases as n decreases. Thus Y_n is nonsingular for $n \le N$ and the roles of X_n and Y_n can be reversed for the identity

$$X_n = Y_n[I - S_{n,N}(Y)] \quad \text{for } n \le N, \tag{4.3}$$

since the Q in this case is $\{Y_n, X_n\} = -\{X_n, Y_n\}' = -(-I)' = I$. Compare (4.2) and (4.3) for the identity

$$I = (I - S_{n,N}(Y))(I + S_{n,N}(X)), \qquad n \le N. \tag{4.4}$$

Since the second factor is increasing as n decreases and exceeds I for $n < N$, the first factor must be decreasing as n decreases and satisfy $0 < I - S_{n,N}(Y) < I$ for $n < N$. Thus $S_{n,N}(Y)$ is bounded above, Hermitian and increasing as n decreases; hence it has a positive-definite limit $S_{-\infty,N}(Y)$. (The solution Y is dominant at $-\infty$.) The *recessive solution* at $-\infty$ is formed as in [23], [2], and [6] (for the recessive solution at ∞) by defining a new solution by

$$Z_n = Y_n[S_{-\infty,N}(Y) - S_{n,N}(Y)], \qquad n \le N. \tag{4.5}$$

Then $Y_n^{-1}Z_n \to 0$ as $n \to -\infty$. Also, Z_n is nonsingular for $n \leq N$ and Z_n is prepared. Reverse the roles for

$$Y_n = Z_n[(S_{-\infty,N}(Y))^{-1} + S_{n,N}(Z)]. \tag{4.6}$$

Let μ be the maximum eigenvalue of $(S_{-\infty,N}(Y))^{-1}$. Then $Z_n^{-1}Y_n \leq \mu I + S_{n,N}(Z)$ and all eigenvalues of $S_{n,N}(Z)$ must go to ∞ as $n \to -\infty$. Thus

$$[S_{n,N}(Z)]^{-1} \to 0 \quad \text{as } n \to -\infty. \tag{4.7}$$

THEOREM 4.1. *Suppose A_n and C_n are real symmetric with C_n nonsingular for all n. Assume that* (1.1) *is disconjugate on* $(-\infty, N+1]$. *Then there exists a solution Z_n of* (1.2) *which is recessive at $-\infty$. It has the following properties*:

(i) *Z_n is nonsingular for $n \leq N$*;

(ii) *Z_n is prepared*;

(iii) *$Z_n'C_{n-1}Z_{n-1} > 0$ for $n < N$*;

(iv) *$[S_{n,N}(Z)]^{-1} \to 0$ as $n \to -\infty$*;

(v) *If Y_n is any prepared solution with $\{Y_n, Z_n\}$ nonsingular, then Y_n is nonsingular for n near $-\infty$ and $Y_n^{-1}Z_n \to 0$, as $n \to -\infty$.*

The proof of condition (v) is immediate from the proof of part (iii) of Theorem 4.1 of [6], given on pp. 19–20.

5. A Reid-type construction of the recessive solution at $-\infty$. Assume that (1.1) is disconjugate on $(-\infty, N+1]$. For $M < N$, let $U_n(M,N)$ denote the solution of (1.2) which satisfies the boundary conditions

$$U_M(M,N) = 0, \qquad U_N(M,N) = I. \tag{5.1}$$

Assume that Z_n is constructed as in the previous section. Without loss of generality, assume $Z_N = I$. Then there exists a matrix Q_M such that

$$U_n(M,N) = Z_n(I - S_{n,N}(Z)Q_M), \qquad n \leq N. \tag{5.2}$$

In particular, $n = M$ gives

$$Q_M = [S_{M,N}(Z)]^{-1} \to 0 \quad \text{as } M \to -\infty \tag{5.3}$$

and hence

$$Z_n = \lim_{M \to -\infty} U_n(M,N) \quad \text{for } n \leq N. \tag{5.4}$$

Thus there can be only one solution Z_n with $Z_N = I$ and property (iv) of Theorem 4.1.

Reid [36] used variational methods to establish that the recessive ("principal") solution at ∞ of a self-adjoint disconjugate linear matrix differential equation was generated as a limiting case of a two-point boundary-value problem. The author, while a student under Professor Reid, once asked him how he knew to use that construction. He replied that he was motivated by Sansone's discussion of the Thomas–Fermi equation. See Sansone [39, pp. 445–450] for a summary of the literature and a proof of existence and uniqueness of a solution of the boundary-value problem

$$x^{1/2}y'' = y^{3/2}, \quad y(0) = 1, \quad \lim_{x \to +\infty} y(x) = 0.$$

Reid's variational proof is patterned after the construction of Sansone. However, as done here, it requires less variational understanding to use Hartman's construction for existence and Reid's construction for uniqueness.

Gautschi [19] credits the analogous backward recurrence construction for the recessive solution of scalar difference equations to J.C.P. Miller. This method was also used by Olver and Sookne [29].

THEOREM 5.1. *Suppose A_n and C_n are real symmetric, C_n is nonsingular, and* (1.1) *is disconjugate on* $(-\infty, N+1]$. *Then the recessive solution Z_n of* (1.2) *with $Z_N = I$ is given by* (5.4) *for $U_n(M,N)$ the unique solution of* (1.2) *which satisfies the boundary conditions* (5.1).

6. Definition, properties, and construction of W_n^+.

Assume that (1.1) is disconjugate on $(-\infty, N+1]$ and Z_n is the recessive solution at $-\infty$ with $Z_N = I$. Then

$$Z_n' C_{n-1} Z_{n-1} > 0 \quad \text{for } n \le N, \tag{6.1}$$

and W_n^+ may be defined by

$$W_n^+ = (C_{n-1}\Delta Z_{n-1}) Z_{n-1}^{-1} \quad \text{for } n \le N+1. \tag{6.2}$$

Then

$$\begin{aligned} W_n^+ + C_{n-1} &= C_{n-1} Z_n Z_{n-1}^{-1} \\ &= (Z_{n-1}^{-1})'(Z_{n-1}' C_{n-1} Z_n) Z_{n-1}^{-1} > 0 \end{aligned} \tag{6.3}$$

for $n \le N$ and W_n^+ is a real symmetric solution of the Riccati equation

$$W_{n+1} = A_n + W_n (W_n + C_{n-1})^{-1} C_{n-1}. \tag{6.4}$$

Suppose that W_n is any other Hermitian solution of (6.4) which is left extensible. We now show that

$$W_n \le W_n^+ \tag{6.5}$$

for n in some neighborhood of $-\infty$. Let N be chosen so that $W_n + C_{n-1}$ is nonsingular for $n \le N$. Furthermore, it is possible to choose N so that $W_n + C_{n-1}$ is positive definite for $n \le N$. This is possible because of the corollary to the Sturm theorem, Theorem 3.2, applied to the solution X_n of (1.2) defined by $X_N = I$ and

$$X_{n-1} = (W_n + C_{n-1})^{-1} C_{n-1} X_n, \qquad n \le N, \tag{6.6}$$

which necessarily has

$$X_{n-1}'(W_n + C_{n-1}) X_{n-1} = X_{n-1}' C_{n-1} X_n > 0 \tag{6.7}$$

for n in some neighborhood of $-\infty$. Then, for N chosen as above, we have

$$Z_n = X_n (I - S_{n,N}(X) Q) \quad \text{for } n \le N. \tag{6.8}$$

For $U_n(M,N)$ as in (5.1), there exists a matrix Γ_M such that

$$U_n(M,N) = X_n (I - S_{n,N}(X)\Gamma_M), \qquad n \le N. \tag{6.9}$$

In particular, $\Gamma_M = (S_{M,N}(X))^{-1}$ and

$$Q = \lim_{M\to-\infty} \Gamma_M = \lim_{M\to-\infty} (S_{M,N}(X))^{-1}, \tag{6.10}$$

where

$$Q = \{X_n, Z_n\} = X'_{n-1}C_{n-1}Z_n - X'_nC_{n-1}Z_{n-1} = C_NZ_{N+1} - X'_{N+1}C_N \tag{6.11}$$

by using $n = N + 1$. Since we wish to compare W_n to W_n^+, note that

$$W_n^+ - W_n = C_{n-1}Z_nZ_{n-1}^{-1} - C_{n-1}X_nX_{n-1}^{-1}, \qquad n \le M + 1. \tag{6.12}$$

But $n = N + 1$ and $X'_{N+1}C_NX_N = X'_NC_NX_{N+1}$ give

$$W_{N+1}^+ - W_{N+1} = Q = \lim_{M\to-\infty} (S_{M,N}(X))^{-1} \ge 0 \tag{6.13}$$

with equality if and only if X is recessive at $-\infty$. Since N can be chosen arbitrarily close to $-\infty$ we have the "maximality" of W_n^+ stated in the following theorem.

THEOREM 6.1. *Suppose A_n and C_n are real symmetric, C_n is nonsingular and* (1.1) *is disconjugate on* $(-\infty, N + 1]$. *Then W_n^+, the solution of the Riccati equation* (6.4) *corresponding to any solution Z of* (1.2) *which is recessive at $-\infty$ has representation*

$$W_n^+ = \lim_{M\to-\infty} C_{n-1}(\Delta U_{n-1}(M, N))U_{n-1}^{-1}(M, N) \tag{6.14}$$

for $n \le N$. This solution W_n^+ has the "primordial maximality" property: For any left extensible Hermitian solution W_n of (6.4), *the inequality*

$$W_n \le W_n^+ \tag{6.15}$$

is satisfied on some neighborhood of $-\infty$.

7. Continued fraction representations of W_n^+.

The result of Theorem 6.1 could be restated in terms of any solution $U_n(M)$ with

$$U_M(M) = 0, \qquad U_{M+1}(M) \text{ nonsingular.} \tag{7.1}$$

Then $W_n(M)$, defined for $M + 2 \le n \le N$ by

$$W_n(M) = (C_{n-1}\Delta U_{n-1}(M))U_{n-1}^{-1}(M), \tag{7.2}$$

satisfies the Riccati equation

$$W_{n+1}(M) = A_n + W_n(M)[W_n(M) + C_{n-1}]^{-1}C_{n-1}, \qquad n = M + 2, \dots, N - 1. \tag{7.3}$$

If we wish to approximate W_N^+, then we need to know $W_{M+2}(M)$ in order to start the iteration (7.3). Rewrite (1.2) as

$$\begin{aligned} 0 &= -\Delta(C_{n-1}\Delta U_{n-1}(M)) + A_nU_n(M) \\ &= [-(C_n\Delta U_n)U_n^{-1} + (C_{n-1}\Delta U_{n-1})U_n^{-1} + A_n]U_n \\ &= [-W_{n+1}(M) + C_{n-1}(I - U_{n-1}U_n^{-1}) + A_n]U_n \end{aligned}$$

and choose n as $M+1$, for the starting value

$$W_{M+2}(M) = A_{M+1} + C_M \tag{7.4}$$

for the iteration (7.3).

THEOREM 7.1. *Suppose that N is such that* (1.1) *is disconjugate on* $(-\infty, N+1]$. *Then, for any positive real number ϵ, there exists an M in* $(-\infty, N-1)$ *such that the value of $W_N(M)$ defined by the iteration* (7.4), (7.3) *has*

$$||W_N^+ - W_N(M)||_2 < \epsilon. \tag{7.5}$$

This shows that computation of W_n^+ requires knowledge of all "previous" coefficients, namely, A_{n-1}, C_{n-2}, A_{n-2}, C_{n-3}, ... If the coefficients are periodic, then these computations are possible; results for the K-periodic case are given in §9.

8. Approximants as iterates of linear fractional transformations.

Define matrix linear fractional transformations T_n by

$$T_n(W) = A_n + W(W + C_{n-1})^{-1}C_{n-1} \ . \tag{8.1}$$

The approximant $W_N(M)$ of W_N^+ given in Theorem 7.1 can then be expressed in terms of these linear fractional transformations. Recall the iteration

$$W_{M+2}(M) = A_{M+1} + C_M \ , \tag{8.2}$$

$$W_{n+1}(M) = A_n + W_n(M)[W_n(M) + C_{n-1}]^{-1}C_{n-1}, \quad n = M+2, \dots, N-1. \tag{8.3}$$

Then

$$W_N(M) = T_{N-1} \circ T_{N-2} \circ \cdots \circ T_{M+3} \circ T_{M+2}(A_{M+1} + C_M). \tag{8.4}$$

Given an initial Hermitian matrix $\mathcal{Z}$, define the function $\mathcal{T}(\mathcal{Z})$ by

$$\mathcal{T}(\mathcal{Z}) = T_{N-1} \circ T_{N-2} \circ \cdots \circ T_{M+3} \circ T_{M+2}(\mathcal{Z}). \tag{8.5}$$

Then $\mathcal{T}(\mathcal{Z})$ is the result of starting the solution of the Riccati recurrence with $W_{M+2} = \mathcal{Z}$ and following it to the right to W_N. It will be shown that this expression can be written in terms of a linear fractional transformation involving $\mathcal{Z}$ and two solutions of the linear equation. These "analytic" methods are being developed for the treatment of the case of periodic coefficients of the next section.

The prototype for this construction is that of Wall [45, pp. 13–16]. We first summarize the relevant points of his definition of continued fractions as compositions of Möbius transformations.

Wall defines t_p by

$$t_0(w) = b_0 + w, \quad t_p(w) = \frac{a_p}{b_p + w}, \quad p = 1, 2, 3, \dots.$$

The associated continued fraction generated by

$$\lim_{n\to\infty} t_0 t_1 \cdots t_n(0), \quad \text{i.e.,} \quad \lim_{n\to\infty} t_0 t_1 \cdots t_{n+1}(\infty),$$

is

$$b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \ddots}}.$$

The quantity

$$t_0 t_1 \cdots t_n(0) = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \ddots + (a_n/b_n)}}$$

is called the nth *approximant.* The 0th approximant is $t_0(0) = b_0$.

Composition of the linear fractional transformations gives

$$t_0 t_1 \cdots t_n(w) = \frac{A_{n-1}w + A_n}{B_{n-1}w + B_n},$$

where the quantities A_{n-1}, A_n, B_{n-1}, B_n satisfy the *fundamental recurrence formulas:*

$$\begin{aligned} A_{-1} &= 1, \quad A_0 = b_b, \quad B_{-1} = 0, \quad B_0 = 1; \\ A_{p+1} &= b_{p+1}A_p + a_{p+1}A_{p-1}, \\ B_{p+1} &= b_{p+1}B_p + a_{p+1}B_{p-1}, \end{aligned}$$

for $p = 0, 1, 2, \ldots$.

Finally, Wall inductively computes the determinant of the composite transformation $t(w) = t_0 t_1 \cdots t_n(w)$ by

$$\begin{vmatrix} A_{n-1} & A_n \\ B_{n-1} & B_n \end{vmatrix} = \begin{vmatrix} A_{n-1} & b_n A_{n-1} + a_n A_{n-2} \\ B_{n-1} & b_n B_{n-1} + a_n B_{n-2} \end{vmatrix} = -a_n \begin{vmatrix} A_{n-2} & A_{n-1} \\ B_{n-2} & B_{n-1} \end{vmatrix},$$

so that

$$A_{n-1}B_n - A_n B_{n-1} = (-1)^n a_0 a_1 \cdots a_n, \qquad n = 0, 1, 2, \ldots,$$

where a_0 must be taken equal to unity.

Note that in our case all the a_p are 1, so the determinants are all ±1. Since we are taking only every other approximant, we expect constancy of these determinants as functions of n.

The essential idea which we now emulate is that *the corresponding solution of the linear system gives "projective coordinates" for the approximants.* Given any two linearly independent matrix solutions U_n and V_n of the linear system (1.2), any other solution X_n may be expressed uniquely in terms of U_n and V_n. Corresponding to the initial-value problem

$$W_{M+2} = \mathcal{Z}, \quad W_{n+1} = T_n(W_n), \quad n \geq M + 2,$$

let X_n be defined as the corresponding solution of the linear equation defined by $X_{M+1} = I$ and

$$X_n = C_{n-1}^{-1}(W_n + C_{n-1})X_{n-1}, \qquad M + 2 \leq n \leq N. \tag{8.6}$$

Then $X_{M+2} = C_{M+1}^{-1}\mathcal{Z} + I$. Let U_n and V_n be the solutions of (1.2) which satisfy the initial conditions $U_{M+1} = I = U_{M+2}$ and $V_{M+1} = 0$, $V_{M+2} = C_{M+1}^{-1}$, respectively.

Then U_n and V_n are prepared with $\{U_n, V_n\} = I$. Thus, $X_n = U_n\Gamma_1 + V_n\Gamma_2$ for some constant matrices Γ_1 and Γ_2, namely, $\Gamma_1 = I$ and $\Gamma_2 = \mathcal{Z}$, i.e.,

$$X_n = U_n + V_n\mathcal{Z}. \tag{8.7}$$

Then

$$W_n = \{C_{n-1}(\Delta V_{n-1})\mathcal{Z} + C_{n-1}\Delta U_{n-1}\}\{V_{n-1}\mathcal{Z} + U_{n-1}\}^{-1}, \tag{8.8}$$

and we have shown the following.

THEOREM 8.1. *For N larger than $M+2$, the composite transformation $\mathcal{T}(\mathcal{Z})$ defined by* (8.5)*, which carries the solution starting at $W_{M+2} = \mathcal{Z}$ to $W_N = \mathcal{T}(\mathcal{Z})$, is the matrix linear fractional transformation of $\mathcal{Z}$ given by*

$$\mathcal{T}(\mathcal{Z}) = \{E\mathcal{Z} + F\}\{G\mathcal{Z} + H\}^{-1}, \tag{8.9}$$

where

$$\begin{aligned} &E = C_{N-1}\Delta V_{N-1}, \qquad F = C_{N-1}\Delta U_{N-1}, \\ &G = V_{N-1}, \qquad H = U_{N-1} \end{aligned} \tag{8.10}$$

for solutions U_n and V_n of (1.2) *with initial conditions*

$$U_{M+1} = I = U_{M+2} \quad \text{and} \quad V_{M+1} = 0, \quad V_{M+2} = C_{M+1}^{-1}, \tag{8.11}$$

respectively. Furthermore, the "determinant" of this transformation is

$$\begin{vmatrix} E & F \\ G & H \end{vmatrix} = E'H - G'F = I, \tag{8.12}$$

and we also have the remaining "symplectic" conditions

$$E'G - G'E = 0 \quad \text{and} \quad F'H - H'F = 0. \tag{8.13}$$

The concepts of symplectic matrices and associated linear fractional transformations will be discussed in §13.

9. K-periodic coefficients.

Suppose (1.1) is disconjugate on $(-\infty,\infty)$. Assume that the coefficients are K-periodic, i.e., for some positive integer K,

$$A_{n+K} = A_n, \quad C_{n+K} = C_n \qquad \text{for all } n. \tag{9.1}$$

Then as in [6, p. 2] or [37] maximality implies that $W_n^+ \geq W_{n\pm K}^+$ and hence W_n^+ is K-periodic. The iteration of Theorem 7.1 can be continued until convergence with only a finite number of coefficients. Indeed, set $M+2 = N-K$ in (7.4) and iterate by setting $W_{M+2} = W_N$.

THEOREM 9.1. *Suppose that* (1.1) *is disconjugate on $(-\infty,\infty)$ and K-periodic. In the following algorithm, W_N converges to W_N^+.*

(1) *Initialize by setting $W_{N-K} = A_{N-1} + C_{N-2}$. Then iterate by the following.*

(2) *For $n = N-K,\ldots,N-1$, set $W_{n+1} = A_n + W_n[W_n + C_{n-1}]^{-1}C_{n-1}$.*

(3) *Compare W_{N-K} with W_N. If they "agree," stop; else, set $W_{N-K} = W_N$ and return to* (2)*.*

Theorem 8.1 allows us to step from W_{N-K} to W_N by using the composite function $\mathcal{T}$. Then the problem is reduced to finding a fixed point for the linear fractional transformation $\mathcal{T}(\mathcal{Z})$. Other values of W_n can then be found by iterating the Riccati equation. This role of the linear equation is somewhat like the corresponding Floquet theory in linear ordinary differential equations with periodic coefficients.

THEOREM 9.2. *The algorithm of Theorem* 9.1 *can be replaced by the following: As above, set* $M+2=N-K$.

(1) *Initialize by setting* $\mathcal{Z}=A_{N-1}+C_{N-2}$.

(2) *For* $n=M+2,\ldots,N$ *compute the solutions* U_n *and* V_n *of* (1.2) *which satisfy the initial conditions* (8.11).

(3) *Compute* E, F, G, *and* H *by* (8.10).

(4) *Iterate by replacing* $\mathcal{Z}$ *by* $\mathcal{T}(\mathcal{Z})$ *of* (8.9) *until convergence.*

(5) *Replace* W_{N-K} *by* $\mathcal{Z}$.

(6) *For* $n=N-K,\ldots,N-1$, *set* $W_{n+1}=A_n+W_n[W_n+C_{n-1}]^{-1}C_{n-1}$.

The fact that, for our setting, a convergent periodic continued fraction converges to a solution of a quadratic equation

$$\mathcal{Z}\{G\mathcal{Z}+H\}=E\mathcal{Z}+F \tag{9.2}$$

is a generalization of Theorem 176 of Hardy [22]. (The converse of that result, Theorem 177 of Hardy [22], says that the continued fraction which represents a quadratic surd is periodic. According to Hardy [22, p. 152] that result is "Lagrange's most famous contribution to the theory.")

The special case of $K=1$ is the constant coefficient case.

COROLLARY 9.1. *Suppose* A_n *and* C_n *are constant and* (1.1) *is disconjugate on* $(-\infty,\infty)$. *Then* W_n^+ *is constant and satisfies the discrete algebraic Riccati equation*

$$W=A+W[W+C]^{-1}C. \tag{9.3}$$

Initialize by $W=A+C$ *and iterate by overwriting* W *with* $A+W[W+C]^{-1}C$ *repeatedly in order to obtain* W^+. *Furthermore, for any choice of the integer* K, *the algorithm of Theorem* 9.2 *can also be used on the constant coefficient case.*

We have shown that the periodic case can also be solved by reducing the problem to the solving of a matrix quadratic equation (9.2).

For noniterative solution methods for matrix quadratic equations of Hamiltonian form, see the papers of Potter [34], Vaughan [44], Hewer [24], Laub [26], and Byers [9], as well as the references contained therein.

10. Positive definite coefficients.

One case where we know that (1.1) is disconjugate on $(-\infty,\infty)$ is the case where

$$A_n>0,\qquad C_n>0\quad\text{for all } n. \tag{10.1}$$

Indeed, condition (vi) of Theorem 3.1 is immediate. (The tridiagonal matrix of the Corollary to Theorem 3.1 has a property which can be thought of as a generalization of diagonal dominance.) In this case the approximants $W_N(M)$ of Theorem 7.1 are positive definite and satisfy

$$W_{n+1}(M)=A_n+[C_{n-1}^{-1}+W_n^{-1}(M)]^{-1},\qquad n=M+2,\ldots,N-1. \tag{10.2}$$

This is because (7.4) makes $W_{M+2}(M) > 0$ and (7.3) inductively makes $W_{n+1}(M)$ positive definite if $W_n(M)$ is positive definite. Set $D_n = C_n^{-1}$ for the iteration:

$$\begin{aligned} &(10.3) \qquad W_{M+2}(M) = A_{M+1} + [D_M]^{-1} = A_{M+1} + \frac{I}{D_M}, \\ &\qquad\qquad W_{n+1}(M) = A_n + \frac{I}{D_{n-1} + \frac{I}{W_n(M)}}, \qquad n = M+2, \dots, N-1. \end{aligned}$$

Thus, in modern continued fraction notation,

$$(10.4) \qquad W_N(M) = A_{N-1} + \frac{I}{D_{N-2}\,+}\,\frac{I}{A_{N-2}\,+}\,\frac{I}{D_{N-3}\,+}\,\cdots\,+\,\frac{I}{A_{M+1}\,+}\,\frac{I}{D_M}$$

for $M + 2 \leq N$.

THEOREM 10.1. *Suppose A_n and C_n are positive definite on $(-\infty, \infty)$. Then* (1.1) *is disconjugate on $(-\infty, \infty)$ and for any n the approximants* ($D_n \equiv C_n^{-1}$)

$$(10.5) \qquad W_n(M) = A_{n-1} + \frac{I}{D_{n-2}\,+}\,\frac{I}{A_{n-2}\,+}\,\cdots\,+\,\frac{I}{A_{M+1}\,+}\,\frac{I}{D_M}$$

converge monotonically from above to W_n^+ as $M \to -\infty$. Furthermore, the approximants

$$(10.6) \qquad D_{n-1} + \frac{I}{A_n\,+}\,\frac{I}{D_n\,+}\,\frac{I}{A_{n+1}\,+}\,\cdots\,+\,\frac{I}{A_{N-1}\,+}\,\frac{I}{D_{N-1}}$$

converge monotonically from below to $-(W_n^-)^{-1}$ as $N \to \infty$, where W_n^- is the unique minimal solution of (1.5).

The unique, eventually minimal solution W_n^- corresponds to the recessive solution at ∞. It follows readily from Theorem 4.1 of [6] that the approximants are generated by

$$(10.7) \qquad W_n(N) = (C_{n-1}\Delta U_{n-1}(N))U_{n-1}^{-1}(N), \qquad n \leq N,$$

where $U_n(N)$ satisfies $U_N(N) = 0$ and $U_{N-1}(N)$ nonsingular. Then

$$W_N(N) = C_{N-1}(0 - U_{N-1}(N))U_{N-1}^{-1}(N) = -C_{N-1}$$

and

$$W_{n+1}(N) = A_n + W_n(N)[W_n(N) + C_{n-1}]^{-1}C_{n-1}, \qquad n < N.$$

Inducting to the left, $W_{n+1}(N)$ negative definite implies that

$$0 > W_{n+1}(N) - A_n = W_n(N)[W_n(N) + C_{n-1}]^{-1}C_{n-1}, \qquad n < N.$$

Hence, if $W_{n+1}(N)$ is negative definite, then $W_n(N)$ is nonsingular and satisfies

$$W_{n+1}(N) - A_n = [C_{n-1}^{-1} + W_n^{-1}(N)]^{-1} < 0.$$

Thus, for $n \leq N$,

$$W_n^{-1}(N) + D_{n-1} = -[A_n - W_{n+1}(N)]^{-1},$$

$W_n(N)$ is negative definite, and

$$-W_n^{-1}(N) = D_{n-1} + \frac{I}{A_n - W_{n+1}(N)} \tag{10.8}$$

for $n = N-1, N-2, \ldots$. Since $-W_N(N) = D_{N-1}^{-1}$, relation (10.6) follows by iterating (10.8). Convergence to $(W_n^-)^{-1}$ follows since the approximants $W_n(N)$, i.e.,

$$W_n(N) = -\left[D_{n-1} + \frac{I}{A_n} + \frac{I}{D_n} + \cdots + \frac{I}{D_{N-1}}\right]^{-1}$$

are known to converge to W_n^- as $N \to \infty$ [6, Thm. 4.1, part (v)].

COROLLARY 10.1. *Suppose A_n and C_n are positive definite and K-periodic. Then the iteration which starts with $W_n = A_{n-1} + C_{n-2}$ and iterates by*

$$W_n = A_{n-1} + \frac{I}{D_{n-2}} + \frac{I}{A_{n-2}} + \cdots + \frac{I}{A_{n-K}} + \frac{I}{D_{n-K-1}} + \frac{I}{W_n} \tag{10.9}$$

converges montonically from above to W_n^+. Convergence may be accelerated by using Theorem 9.2. The corresponding iteration which starts with $V_n = D_{n-1}$ and iterates by

$$V_n = D_{n-1} + \frac{I}{A_n} + \frac{I}{D_n} + \cdots + \frac{I}{D_{n+K-2}} + \frac{I}{A_{n+K-1}} + \frac{I}{V_n} \tag{10.10}$$

converges monotonically from below to $-(W_n^-)^{-1}$.

Note that (10.10) could be rewritten as

$$V_n = D_{n-K-1} + \frac{I}{A_{n-K}} + \frac{I}{D_{n-K}} + \cdots + \frac{I}{D_{n-2}} + \frac{I}{A_{n-1}} + \frac{I}{V_n}, \tag{10.11}$$

which generates the "reverse" periodic continued fraction to (10.9). Hence the above corollary provides a matrix interpretation of the scalar result of Galois [18] cited in [28, Cor. 1, p. 26]. Indeed, the results of §9 establish that for each n, there exists a matrix quadratic equation for which W_n^+ is a solution.

It is important to *not* write the continued fraction representation of W_n^+ of Theorem 10.1 as

$$W_n^+ = A_{n-1} + \frac{I}{D_{n-2}} + \frac{I}{A_{n-2}} + \frac{I}{D_{n-3}} + \cdots$$

since W_n^+ is obtained as a contraction (Perron [31, Chap. I, §4, pp. 10–16]) of this continued fraction. Indeed, these continued fractions can diverge even though their even and odd approximants converge. Wall has the following result [45, p. 28, Thm. 6.1]: If the series $\sum |b_p|$ converges, then the continued fraction

$$\frac{1}{b_1} + \frac{1}{b_2} + \frac{1}{b_3} + \cdots$$

diverges. The sequences of its even and odd numerators and denominators, A_{2p}, A_{2p+1}, B_{2p}, B_{2p+1}, converge to finite limits F_0, F_1, G_0, G_1, respectively, where

$$F_1G_0 - F_0G_1 = 1.$$

11. Positive-definite constant coefficients. In the case where A and C are positive-definite constant matrices, W_n^+ and W_n^- are of period 1, hence constant. The existence and uniqueness of W^+ and the relationship with W^- were obtained in [6, Thm. 8.1] by the method of time reversal. We summarize the known facts in light of the above continued fraction representations.

THEOREM 11.1. *Suppose A and C are positive definite. Then for $D = C^{-1}$,*

$$W^+ = \lim \quad A + \frac{I}{D} + \frac{I}{A} + \frac{I}{D} + \cdots + \frac{I}{D} \tag{11.1}$$

and

$$-W^- = \lim \quad \frac{I}{D} + \frac{I}{A} + \frac{I}{D} + \cdots + \frac{I}{D}. \tag{11.2}$$

In particular, $W^+ = A - W^-$. Furthermore, if W is any Hermitian solution of

$$W = A + W[W+C]^{-1}C, \quad i.e., \ WC^{-1}W - AC^{-1}W - A = 0, \tag{11.3}$$

then

$$W^- \leq W \leq W^+.$$

Also W^+ and W^- are the only solutions which are positive definite and negative definite, respectively.

The estimates $W^- > -C$ and $A < W^+ < A + C$ obtained in [6] are immediate from these continued fraction representations.

The two forms in (11.3) are equivalent since both are equivalent to

$$(W - A)C^{-1}(W + C) = W. \tag{11.4}$$

Indeed, if W is a solution of (11.4) and u is a vector with $(W + C)u = 0$, then $0 = (W - A)C^{-1}(W + C)u = Wu = -Cu$ and therefore $u = 0$. Thus, any solution W of (11.4) has $W + C$ nonsingular.

An example where additional Hermitian solutions of (11.3) exist is obtained by taking A and C as scalar matrices $A = aI_n$ and $C = cI_n$, with a and c positive. For w^+ and w^-, the positive and negative roots of

$$w^2 - aw - ac = 0,$$

we have $W^+ = w^+ I_n$ and $W^- = w^- I_n$. But (11.3) and (11.4) have 2^n solutions consisting of diagonal matrices W with diagonal entries chosen from w^+ and w^-.

12. Numerical methods. Numerical experiments with Matlab show that convergence of these continued fraction representations can be quite slow, although they are stable and preserve symmetry. In the variable coefficient case, one is faced with the difficult choice of how many "terms" are necessary. The continued fraction must be computed from "tail to head." A second try with more terms gives a feeling for how many significant digits have been achieved. If more terms are needed for a good estimate, then a completely new calculation must be carried out. This does give an idea of the rate of convergence and some feel for how far one is from convergence to machine precision. This difficulty of knowing how many terms and successively recomputing was addressed in the scalar case by Gautschi's "second" and "third" algorithms; see [19, §4, pp. 42–46].

Attention is now focused on the constant coefficient case. A few terms of the continued fraction representations of Theorem 10.1 can be used to provide starting values for the more rapidly convergent Newton's method. Consider the matrix function F defined by

$$F(W) = WDW - ADW - A, \tag{12.1}$$

where $D = C^{-1}$ and A are positive definite. Then

$$F(W + H) = F(W) + WDH + HDW - ADH + HDH. \tag{12.2}$$

If we neglect the "quadratic" term HDH and choose H so that $F(W + H) = 0$, then Newton's method is to replace W by $W + H$, where H satisfies

$$(A - W)DH - HDW = F(W). \tag{12.3}$$

If A and D are positive definite and $W > A$, then $A - W$ and W have no eigenvalues in common and (12.3) has exactly one solution H; furthermore, H is real symmetric. Indeed, use the numerically stable Cholesky decomposition of D, $D = LL'$, for L lower triangular with positive diagonal entries [20, p. 89] in order to rewrite (12.3) as

$$L'(A - W)LX - XL'WL = L'F(W)L$$

for $X = L'HL$ in order to obtain an equation of the form

$$XP + QX = R, \tag{12.4}$$

where P and Q are the negative-definite matrices $P = -L'WL$, $Q = L'(A - W)L$, and $R = L'F(W)L$. If we choose $q = (||P||_2 + ||Q||_2)/2$ in Smith's Algorithm [43], then (12.4) can be written in the form

$$(qI - Q)X(qI - P) - (qI + Q)X(qI + P) = -2qR. \tag{12.5}$$

Rewrite this equation as

$$X - UXV = S \tag{12.6}$$

for $U = (qI-Q)^{-1}(qI+Q)$, $V = (qI+P)(qI-P)^{-1}$, $S = -2q(qI-Q)^{-1}R(qI+P)^{-1}$. Then the solution X is given by

$$X = \sum_{k=1}^{\infty} U^{k-1}SV^{k-1}, \tag{12.7}$$

which converges since U and V have 2-norm less than 1 (see Smith [43, p. 199]). The solution X is obtained as the limit of the iteration $Y_0 = S$, $Y_{\nu+1} = Y_\nu + U^{2^\nu}Y_\nu V^{2^\nu}$. That is, Smith speeds up the convergence by repeatedly squaring U and V. Thus X gives H and W is replaced by $W + H$. The iteration can by started with W as any of the approximants of W^+ of (11.1). Then H can be obtained by Smith's algorithm. One step of the iteration (10.2) or Theorem 9.2 can be used to stabilize and symmetrize the iterations in between using the Newton method. It is to be noted that if $D^{1/2}$ were used instead of L, the Newton iteration would be unstable. While it is not obvious from (12.4) that H is symmetric, it is true that this solution H is also a solution of the Newton iteration for

$$G(W) = WDW - \tfrac{1}{2}ADW - \tfrac{1}{2}WDA - A.$$

That is,

$$(\tfrac{1}{2}A - W)DH + HD(\tfrac{1}{2}A - W) = G(W) = F(W).$$

For L as before, rewrite this as

$$L'(\tfrac{1}{2}A - W)DHL + L'HD(\tfrac{1}{2}A - W)L = L'F(W)L,$$

i.e., as

$$PK + KP = R$$

for $K = L'HL$, $P = L'(\frac{1}{2}A - W)L$, and $R = L'F(W)L$. Since P is negative definite if $W > A$, this equation has one and only one solution K, which must be symmetric since K also satisfies

$$K'P + PK' = R.$$

Furthermore, K and H are positive definite if $F(W)$ is positive definite. (See Hale [21, Lemma 1.5, p. 315].)

For A and D positive definite, W^+ must agree with the unique positive-definite solution of the symmetric equation, $G(W) = 0$ (see [6, pp. 33, 34] and Coppel [12]).

13. Pincherle's theorem for symplectic continued fractions.

A famous theorem due to Pincherle [33, Chap. III, §15, pp. 228–230] says that a three term recurrence relation has a recessive solution if and only if the associated continued fraction converges. See Jones and Thron [25, Appen. B] for a statement of this result. Jones and Thron call such a solution *minimal.* Pincherle used the terminology *integrale distinto.* (W. T. Reid later used the term *distinguished* for the corresponding solution of the Riccati differential equation.)

We now construct a theory of symplectic continued fractions in order to investigate the corresponding theorem for the self-adjoint case. We already know that under a disconjugacy hypothesis, we have a recessive solution at ∞, and the corresponding solution of the Riccati equation, W_n^-, has a continued fraction representation. The Fibonacci recurrence, $y_{n+1} = y_n + y_{n-1}$, when written in self-adjoint form

$$-(-1)^n y_{n+1} - (-1)^{n-1} y_{n-1} + (-1)^n y_n = 0, \tag{13.1}$$

where $c_n = (-1)^n$ and $b_n = (-1)^n = a_n$, has the recessive solution $y_n = r_2^n$, for r_2 the negative root of $r^2 = r + 1$, but it fails to be disconjugate on any neighborhood of ∞. Indeed, the sequence $\{c_n r_2^n r_2^{n+1}\}$ changes sign and the Sturm separation theorem (Theorem 3.2) implies that no real solution y_n can have $c_n y_n y_{n+1}$ positive in any neighborhood of ∞. (This can also be seen from the Legendre necessary condition given in [4]. For minimization problems, b_n must be nonnegative and for maximization problems, b_n must be nonpositive. Since $b_n = (-1)^n$, the Jacobi condition cannot be satisfied and every neighborhood of ∞ must contain conjugate intervals.) This example contrasts with differential equations where eventual disconjugacy is equivalent to existence of a recessive solution. Thus, for discrete problems, the question of criteria for existence of recessive solutions is more interesting than in the corresponding continuous theory. The purpose of this section is to investigate the question of extension of Pincherle's theorem to systems. Additional insight is provided for scalar self-adjoint recurrence relations.

As noted in Theorem 8.1, we are led to a class of linear fractional transformations of the form

$$T(Z) = (EZ + F)(GZ + H)^{-1} \tag{13.2}$$

whose coefficients satisfy

$$E'G = G'E, \quad F'H = H'F, \quad E'H - G'F = I. \tag{13.3}$$

Associated with the transformation T are the matrices $\mathcal{M}$ and $\mathcal{J}$ defined by

$$\mathcal{M} = \begin{bmatrix} E & F \\ G & H \end{bmatrix}, \quad \mathcal{J} = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}. \tag{13.4}$$

The matrix $\mathcal{M}$ is called a *symplectic matrix* (Siegel [42]) if conditions (13.3) are satisfied, i.e., $\mathcal{M}'\mathcal{J}\mathcal{M} = \mathcal{J}$. Since this condition is equivalent to $\mathcal{M}^{-1} = -\mathcal{J}\mathcal{M}'\mathcal{J}$, the inverse of the symplectic matrix $\mathcal{M}$ is given by

$$\mathcal{M}^{-1} = \begin{bmatrix} H' & -F' \\ -G' & E' \end{bmatrix}. \tag{13.5}$$

Since $\mathcal{M}^{-1}$ is a two-sided inverse, conditions (13.3) are equivalent to

$$EF' = FE', \quad GH' = HG', \quad EH' - FG' = I. \tag{13.6}$$

Note that the symplectic matrices of order $2r$ form a group under multiplication. Associated with a symplectic matrix $\mathcal{M}$ is the *symplectic transformation* $T_{\mathcal{M}}$. It follows that

$$T_{\mathcal{M}_1} \circ T_{\mathcal{M}_2} = T_{\mathcal{M}_1\mathcal{M}_2}. \tag{13.7}$$

Hence, the symplectic transformations form a transformation group. We also write T_n for the transformation associated with a symplectic matrix $\mathcal{M}_n$. Note that $\mathcal{M}$ and $-\mathcal{M}$ give the same transformation T. (Seigel identifies these matrices.) Symplectic transformations are of special interest not only because the transformation $\mathcal{T}$ of §8 is symplectic, but because the transformations T_n defined by (8.1) are symplectic. Indeed, if we set $D_{n-1} = C_{n-1}^{-1}$, then the transformation

$$T_n(Z) = A_n + Z(Z + C_{n-1})^{-1}C_{n-1} \tag{13.8}$$

can be rewritten as

$$T_n(Z) = A_n + Z[D_{n-1}Z + I]^{-1} = \{A_n(D_{n-1}Z + I) + Z\}\{D_{n-1}Z + I\}^{-1},$$

i.e., in symplectic form,

$$T_n(Z) = (E_nZ + F_n)(G_nZ + H_n)^{-1}, \tag{13.9}$$

with associated symplectic matrix

$$\mathcal{M}_n = \begin{bmatrix} E_n & F_n \\ G_n & H_n \end{bmatrix} = \begin{bmatrix} A_nD_{n-1} + I & A_n \\ D_{n-1} & I \end{bmatrix}. \tag{13.10}$$

However, those transformations run the wrong direction since they were designed to approximate W_n^+. Thus we need to take a fresh start to compare with the usual Pincherle theorem.

Since Pincherle's theorem concerns the recessive solution at ∞, our continued fractions are motivated from the representation

$$(13.11)\quad -W_n^- = \lim_{N\to\infty} \frac{I}{D_{n-1}} + \frac{I}{A_n} + \frac{I}{D_n} + \frac{I}{A_{n+1}} + \cdots + \frac{I}{A_{N-1}} + \frac{I}{D_{N-1}},$$

which is valid when the A_n and D_n are positive definite.

Recall the definition of $\mathcal{T}$ given in (8.5), i.e.,

$$(13.12)\qquad \mathcal{T}(\mathcal{Z}) = T_{N-1} \circ T_{N-2} \circ \cdots \circ T_{M+3} \circ T_{M+2}(\mathcal{Z}).$$

In the statement of Pincherle's theorem given in Jones and Thron [25, p. 403], the continued fraction considered is

$$\frac{a_1}{b_1} + \frac{a_2}{b_2} + \cdots.$$

Thus, we will start with $\mathcal{M}_n$ for $n = 1$. Hence we set $M + 2 = 1$, i.e., $M = -1$, and replace $N - 1$ of (13.12) by N in order to define $\mathcal{T}_N$, for $N \geq 1$, by

$$(13.13)\qquad \mathcal{T}_N(\mathcal{Z}) = T_N \circ T_{N-1} \circ \cdots \circ T_1(\mathcal{Z}).$$

For each positive integer N define a transformation $\mathcal{S}_N$ by

$$\mathcal{S}_N(\mathcal{W}) = \mathcal{Z} \quad \text{provided } \mathcal{W}^{-1} = \mathcal{T}_N(-\mathcal{Z}).$$

Since all the matrices $\mathcal{M}_n$ are symplectic, they are invertible, and the transformation $\mathcal{S}_N$ is remarkably easy to invert, namely, $\mathcal{Z} = \mathcal{S}_N(\mathcal{W}) = -\mathcal{T}_N^{-1}(\mathcal{W}^{-1})$ is given by

$$(13.14)\qquad \mathcal{Z} = -T_1^{-1} \circ T_2^{-1} \circ \cdots \circ T_N^{-1}(\mathcal{W}^{-1}).$$

Let us denote T_n^{-1} by s_n. Then

$$(13.15)\qquad \mathcal{S}_N(\mathcal{W}) = -s_1 \circ s_2 \circ \cdots \circ s_N(\mathcal{W}^{-1}),$$

where, because of (13.5) applied to (13.10), s_n has coefficient matrix

$$(13.16)\qquad \mathcal{M}_n^{-1} = \begin{bmatrix} I & -A_n \\ -D_{n-1} & D_{n-1}A_n + I \end{bmatrix}.$$

Application of Theorem 8.1 with $N - 1$ replaced by N gives the alternative representation of $\mathcal{T}_N$

$$(13.17)\qquad \mathcal{T}_N(\mathcal{Z}) = \{C_N \Delta V_N \mathcal{Z} + C_N \Delta U_N\}\{V_N \mathcal{Z} + U_N\}^{-1},$$

for solutions U_n and V_n of (1.2) with initial conditions

$$(13.18)\qquad U_0 = I = U_1 \quad \text{and} \quad V_0 = 0, \quad V_1 = C_0^{-1},$$

respectively. Since the coefficient matrix is symplectic, we have

$$\mathcal{S}_N(\mathcal{W}) = -\{U_N'\mathcal{W}^{-1} - (C_N\Delta U_N)'\}\{-V_N'\mathcal{W}^{-1} + (C_N\Delta V_N)'\}^{-1},$$

i.e.,

$$\mathcal{S}_N(\mathcal{W}) = \{-(C_N\Delta U_N)'\mathcal{W} + U_N'\}\{-(C_N\Delta V_N)'\mathcal{W} + V_N'\}^{-1}. \tag{13.19}$$

The relevant transformations involving s_n are given by

$$\begin{aligned} s_n(\mathcal{W}^{-1}) &= \{I - A_n\mathcal{W}\}\{-D_{n-1} + (D_{n-1}A_n + I)\mathcal{W}\}^{-1}, \\ s_n(\Theta) &= \{\Theta - A_n\}\{-D_{n-1}\Theta + (D_{n-1}A_n + I)\}^{-1}. \end{aligned} \tag{13.20}$$

We are now prepared to define the continued fraction approximants associated with the sequences of matrices A_n and D_n. We say that $\mathcal{S}_N(0)$ is the Nth *approximant.* For solutions U_n and V_n of (1.2) defined by the initial conditions (13.18) we have the characterization $\mathcal{S}_N(0) = U_N'(V_N')^{-1}$. But each of the T_n and s_n take Hermitian matrices to Hermitian matrices; hence $\mathcal{S}_N(0)$ is Hermitian and

$$\mathcal{S}_N(0) = V_N^{-1}U_N. \tag{13.21}$$

The corresponding *continued fraction* is the sequence of approximants $\{\mathcal{S}_N(0)\}$. We will say that the continued fraction *converges* if V_n is nonsingular for large n and the sequence of approximants has a (finite) matrix limit.

This compares favorably with the relation for $-W_1^-$ given in (13.11) when the coefficients are positive definite. Indeed, in that case

$$\mathcal{S}_1(0) = D_0^{-1} \tag{13.22}$$

and

$$\mathcal{S}_2(0) = \frac{I}{D_0} + \frac{I}{A_1} + \frac{I}{D_1}. \tag{13.23}$$

The iteration in this case, namely, $s_N(\mathcal{W}^{-1})|_{\mathcal{W}=0} = -D_{N-1}^{-1}$ and

$$s_n(\Theta) = -\{D_{n-1} + (A_n - \Theta)^{-1}\}^{-1},$$

allows one to see that (13.15) does yield approximants $\mathcal{S}_N(0)$ corresponding to the right-hand side of (13.11).

The above matrix approach to continued fractions depends heavily on the application of matrix representations of Möbius transformations to scalar continued fractions given by Schwerdtfeger [40], [41].

Another representation of the approximants results from a series interpretation of convergence. The following is related to Theorem 2.1 of Wall [45, pp. 17–18]. However, it is more closely related to the original paper of Pincherle [33, pp. 228–229].

From the bracket function identity, $\{U_n, V_n\} = I$, and relation (2.1) we have

$$U_{n-1}'C_{n-1}V_nV_{n-1}^{-1} - U_n'C_{n-1} = V_{n-1}^{-1}. \tag{13.24}$$

Since V_n is prepared, we may use $C_{n-1}V_nV_{n-1}^{-1} = (V_{n-1}')^{-1}V_n'C_{n-1}$ in (13.24) for the identity

$$U_{n-1}'(V_{n-1}')^{-1}V_n'C_{n-1} - U_n'C_{n-1} = V_{n-1}^{-1}.$$

Therefore,

$$U_{n-1}'(V_{n-1}')^{-1} - U_n'(V_n')^{-1} = V_{n-1}^{-1}C_{n-1}^{-1}(V_n')^{-1}. \tag{13.25}$$

Define R_n by $R_n = (V_{n+1}'C_nV_n)^{-1}$. Then R_n is Hermitian and for $p \geq 1$,

$$\mathcal{S}_{n+p}(0) = \mathcal{S}_n(0) - \sum_{k=0}^{p-1} R_{n+k}. \tag{13.26}$$

Hence, the continued fraction converges if and only if V_n is nonsingular for large n and the series

$$\sum_{k=n}^{\infty} R_k, \quad \text{i.e.,} \quad \sum_{k=n}^{\infty}(V_{k+1}'C_kV_k)^{-1}, \tag{13.27}$$

converges; furthermore, in case the continued fraction converges to a matrix Ω, then Ω is Hermitian and, for n chosen sufficiently large, we have

$$\Omega = V_n^{-1}U_n - \sum_{k=n}^{\infty}(V_{k+1}'C_kV_k)^{-1}. \tag{13.28}$$

These results are summarized as follows.

THEOREM 13.1 (Euler [14]). *Suppose that A_n and C_{n-1} are Hermitian with C_{n-1} nonsingular for $n \geq 1$. Let $D_n = C_n^{-1}$. For U_n and V_n the solutions of* (1.2) *which satisfy the initial conditions*

$$U_0 = I = U_1 \quad \textit{and} \quad V_0 = 0, \quad V_1 = C_0^{-1}, \tag{13.29}$$

assume that V_n is nonsingular for $n \geq M$. Then the approximants $\mathcal{S}_N(0)$, i.e., $V_N^{-1}U_N$, converge if and only if the series $\sum_{k=M}^{\infty}(V_{k+1}'C_kV_k)^{-1}$ converges. Furthermore, if the continued fraction converges to a matrix Ω, then Ω is Hermitian and is given by (13.28) *for $n \geq M$.*

For the history of the equivalence between convergence of continued fractions and convergence of series, see Perron [30, p. 18] and [31, pp. 16–20], and the original paper of Euler [14, p. 63].

We are now prepared to investigate the problem of generalizing Pincherle's theorem. Unfortunately, the development of the theory of recessive solutions given in §§4 and 5 depended upon disconjugacy and the resulting Sturm theorem. A prototype definition for our purposes might be that used by the author in [1, pp. 172–173]. That definition did not use the reduction of order formula. However, we can carry out a reduction of order formula here by using the solution V_n.

We generalize our definition of a recessive solution at ∞ to the following.

A solution Y_n of (1.2) is said to be *recessive at* ∞ if it satisfies the following conditions:

(i) Y_n is a prepared basis;

(ii) If X_n is any solution with $\{X_n, Y_n\}$ nonsingular, then X_n is nonsingular for large n and

$$X_n^{-1}Y_n \to 0 \quad \text{as } n \to \infty. \tag{13.30}$$

Hence, the right-hand counterpart of conclusion (v) of the above Theorem 4.1 is taken as a definition. Also, compare with condition (iv) of Theorem 4.1 of [6, p. 14].

THEOREM 13.2 (Pincherle [33]). *Assume the hypotheses of Theorem* 13.1. *If there exists a recessive solution of* (1.2) *with* Y_0 *nonsingular, then the partial denominators,* V_n*, are nonsingular for large* n *and the continued fraction converges. Conversely, suppose that the denominators* V_n *are nonsingular for large* n *and the continued fraction converges to* Ω*. Then there exists a recessive solution* Y_n *of* (1.2) *with* $Y_0 = I$*. Furthermore,* $\Omega = -C_0\Delta Y_0 = -W_1^-$.

Proof. Assume that (1.2) has a recessive solution Y_n with Y_0 nonsingular. Then, there exist constant matrices P and Q such that

$$Y_n = U_nP + V_nQ \quad \text{for } n = 0, 1, \ldots.$$

Indeed, $P = Y_0$ and $Q = (C_0)(Y_1 - P)$. Also $\{V_n, Y_n\} = -Y_0 = -P$ is nonsingular. Hence V_n is nonsingular for large n,

$$V_n^{-1}Y_nP^{-1} = V_n^{-1}U_n + QP^{-1},$$

and $V_n^{-1}U_n$ has a limit. Hence the continued fraction converges.

Conversely, assume that the continued fraction converges to Ω. Let Y_n be the solution defined by

$$Y_n = U_n - V_n\Omega. \tag{13.31}$$

We will show that Y_n is recessive at ∞. Since V_n is nonsingular for large n, we write

$$V_n^{-1}Y_n = V_n^{-1}U_n - \Omega$$

and observe that $V_n^{-1}Y_n \to 0$ as $n \to \infty$. From the initial conditions $Y_0 = I$, and $Y_1 = I - C_0^{-1}\Omega$, we see that $Y_0'C_0Y_1$ is Hermitian, hence Y_n is a prepared basis. Equation (13.28) gives the identity

$$U_n = V_n\Omega + V_n\sum_{k=n}^{\infty}(V_{k+1}'C_kV_k)^{-1}. \tag{13.32}$$

Substitution of this value of U_n in (13.31) gives

$$Y_n = V_n\sum_{k=n}^{\infty}(V_{k+1}'C_kV_k)^{-1}. \tag{13.33}$$

Now use V_n and Y_n as a basis. Suppose that X_n is a solution with $\{X_n, Y_n\}$ nonsingular. We first establish that there exist constant matrices P and Q such that

$$X_n = V_nP + Y_nQ. \tag{13.34}$$

Indeed, use $n = 0$ and $n = 1$ to conclude that $Q = X_0$ and $P = C_0X_1 + (C_0 - \Omega)X_0$. Then $\{X_1, Y_1\} = -P'$ and P is nonsingular. But the form of Y_n given in (13.33) implies that

$$X_n = V_n\left[P + \sum_{k=n}^{\infty}(V'_{k+1}C_kV_k)^{-1}Q\right]. \tag{13.35}$$

Since the series converges to 0 as $n \to \infty$, we conclude that $V_n^{-1}X_n \to P$ and consequently X_n is nonsingular for large n. Finally,

$$X_n^{-1}Y_n = [V_nP + Y_nQ]^{-1}Y_n = (P + V_n^{-1}Y_nQ)^{-1}V_n^{-1}Y_n \to 0.$$

Thus Y_n is recessive.

The computational value of these results is illustrated by the scalar example of the Fibonacci recurrence (13.1), where the recessive solution is $y_n = r_2^n$, for r_2 the negative root of $r^2 = r + 1$. The solution u_n of the initial-value problem $u_{n+1} = u_{n-1} + u_n$ with $u_0 = 1$, $u_1 = 1$, is the nth Fibonacci number; let us denote it by f_n. Then $v_n = f_{n-1}$ and $c_n = (-1)^n$. Thus (13.28) becomes

$$-w^- = \frac{f_n}{f_{n-1}} - \sum_{k=n}^{\infty}\frac{(-1)^k}{f_{k-1}f_k}.$$

Since the series is alternating with terms decreasing in magnitude to 0, the ratio of successive Fibonacci numbers f_n/f_{n-1} approximates $-w^-$ within $1/(f_{n-1}f_n)$. Furthermore, successive approximants provide a nest for the true value.

We now give the continued fraction representation of W_M^-. Let $U_{M,n}$ and $V_{M,n}$ be the solutions of (1.2) which satisfy the initial conditions

$$U_{M-1} = I = U_M \quad \text{and} \quad V_{M-1} = 0, \quad V_M = C_{M-1}^{-1}. \tag{13.36}$$

Make the definition

$$\mathcal{S}_{M,N}(\mathcal{W}) = -s_M \circ s_{M+1} \circ \cdots \circ s_N(\mathcal{W}^{-1}). \tag{13.37}$$

Then the approximants satisfy

$$\mathcal{S}_{M,N}(0) = V_{M,N}^{-1}U_{M,N} \tag{13.38}$$

and the following corollary gives a continued fraction representation of W_M^-.

COROLLARY 13.1. *Suppose that A_n and C_{n-1} are Hermitian with C_{n-1} nonsingular for $n \geq M$. If* (1.2) *has a recessive solution Y_n at ∞ with Y_{M-1} nonsingular, then the approximants $\mathcal{S}_{M,N}(0)$ are defined for large N and converge to $-W_M^-$.*

The above result extends Theorem 1.1 of Gautschi [19, p. 31]; see also part B of Theorem B.4 of Jones and Thron [25, p. 403]. Gautschi used continued fractions to estimate the Bessel functions $J_n(x)$, which (for fixed x) are recessive solutions of a self-adjoint three term recurrence relation. By doing so, he avoided the inherit instability in calculation of the recessive solution from an initial-value problem for a three term recurrence relation.

Finally, because eventual disconjugacy implies the existence of a recessive solution [6, Thm. 4.1], we have the following sufficient condition for convergence.

COROLLARY 13.2. *Suppose that A_n and C_{n-1} are Hermitian with C_{n-1} nonsingular for $n \geq M-2$. Assume that* (1.2) *is disconjugate on* $[M-2,\infty)$. *Then* (1.2) *has a recessive solution Y_n at ∞ with Y_n nonsingular for $n \geq M-1$, and the approximants $\mathcal{S}_{M,N}(0)$ are defined for $N \geq M$ and converge to $-W_M^-$.*

Note added September 22, 1992. In the time since this paper was written continued fraction representations have been obtained for more general Riccati equations than those considered here. Those equations are variable coefficient versions of those of Vaughan [44]. The paper by the author which contains those results is entitled "Geometric, analytic, and arithmetic aspects of symplectic continued fractions." That paper is to appear in a special volume entitled *Analysis, Geometry and Groups: A Riemann Legacy Volume* edited by T. M. Rassias and H. M. Srivastava, and published by Hadronic Press, Palm Harbor, Florida.

Acknowledgment. The author wishes to express his thanks to the referee for constructive comments and for pointing out the work of MacNerney [27]. The author wishes to thank his colleague, L. J. Lange, for sharing his knowledge of the literature on the subject of continued fractions.

REFERENCES

[1] C. D. AHLBRANDT, *Principal and antiprincipal solutions of selfadjoint differential systems and their reciprocals*, Rocky Mountain J. Math., 2 (1972), pp. 169–182.

[2] C. D. AHLBRANDT AND J. W. HOOKER, *Riccati transformations and principal solutions of discrete linear systems*, in Proc. 1984 Workshop Spectral Theory of Sturm–Liouville Differential Operators, H. G. Kaper and A. Zettl, eds., ANL-84-87, Argonne National Laboratories, Argonne, IL, 1984, pp. 1–11.

[3] ———, *Disconjugacy criteria for second order linear difference equations*, in Proc. International Conference on Qualitative Theory of Differential Equations, W. Allegretto and G.J. Butler, eds., Edmonton, Alberta, Canada, 1984, pp. 15–26.

[4] ———, *A variational view of nonoscillation theory for linear difference equations*, in Proc. Thirteenth Midwest Differential Equations Conference, J. L. Henderson, ed., Institute of Applied Mathematics, University of Missouri-Rolla, Rolla, MO, 1985, pp. 1–21.

[5] ———, *Riccati matrix difference equations and disconjugacy of discrete linear systems*, SIAM J. Math. Anal., 19 (1988), pp. 1183–1197.

[6] ———, *Recessive solutions of symmetric three term recurrence relations*, in Canadian Mathematical Society, Conference Proceedings, 8 (1987), pp. 3–42.

[7] F. M. ARSCOTT, R. LACROIX, AND W. T. SHYMANSKI, *A three-term recursion and the computation of Mathieu functions*, Proc. Eighth Manitoba Conference on Numerical Math. and Computing, Univ. Manitoba, Winnipeg, Manitoba, Canada, 1978, Congress. Numer., XXII, Utilitas Math., Winnipeg, Manitoba, 1979, pp. 107–115.

[8] F. M. ARSCOTT, *A Riccati-type transformation of linear difference equations*, Congr. Numer., 30 (1981), pp. 197–202.

[9] R. BYERS, *Solving the algebraic Riccati equation with the matrix sign function*, Linear Algebra Appl., 85 (1987), pp. 267–279.

[10] S. CHEN AND L. H. ERBE, *Oscillation and nonoscillation for systems of self-adjoint second-order difference equations*, SIAM J. Math. Anal., 20 (1989), pp. 939–949.

[11] S. C. COOPER, *δ-fraction solutions to Riccati Equations*, Analytic Theory of Continued Fractions III, (Redstone CO, 1988), Lecture Notes in Math. 1406, Springer-Verlag, Berlin, 1989, pp. 1–17.

[12] W. A. COPPEL, *Matrix quadratic equations*, Bull. Austral. Math. Soc., 10 (1974), pp. 377-401.

[13] H. DENK AND M. RIEDERLE, *A generalization of a theorem of Pringsheim*, J. Approximation Theory, 35 (1982), pp. 355–363.

[14] L. EULER, *Specimen algorithmi singularis*, Novi Commentarii Academiae Scientiarum Imperialis Petropolitanea, 9 (1762), summary on pp. 10–13; full article on pp. 53–69.

[15] W. FAIR, *Noncommutative continued fractions*, SIAM J. Math. Anal., 2 (1971), pp. 226–232.

[16] ———, *A convergence theorem for noncommutative continued fractions*, J. Approx. Theory, 5 (1972), pp. 74–76.

[17] ———, *Continued fraction solution to the Riccati equation in a Banach algebra*, J. Math. Anal. Appl., 39 (1972), pp. 318–323.

[18] É. GALOIS, *Démonstration d'un théorème sur les fractions continues périodiques*, Annales de Mathématiques de M. Gergonne (Annales de Mathématiques Pures et Appliquées), 19 (1828-29), pp. 294–301; see also Oeuvres Mathématiques d'Évariste Galois, Gauthier-Villars, Paris, 1951, pp. 1–8, and Écrits et Mémoires Mathématiques d'Évariste Galois, by R. Bourgne et J.-P. Azra, Gauthier-Villars, Paris, 1962, pp. 365–378.

[19] W. GAUTSCHI, *Computational aspects of three-term recurrence relations*, SIAM Rev., 9 (1967), pp. 24–82.

[20] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.

[21] J. K. HALE, *Ordinary Differential Equations*, Krieger Publishing Company, Huntington, NY, 1980.

[22] G. H. HARDY, *An Introduction to the Theory of Numbers*, Oxford University Press, London, 1938.

[23] P. HARTMAN, *Self-adjoint non-oscillatory systems of ordinary, second order, linear differential equations*, Duke Math. J., 24 (1957), pp. 25–36.

[24] G. A. HEWER, *Periodicity, detectability and the matrix Riccati equation*, SIAM J. Control, 13 (1975), pp. 1235–1251.

[25] W. B. JONES AND W. J. THRON, *Continued Fractions: Analytic Theorey and Applications*, Encyclopedia of Math. and Its Applications, Volume 11, Addison-Wesley, Reading, MA, 1980.

[26] A. J. LAUB, *A Schur method for solving algebraic Riccati equations*, IEEE Trans. Automatic Control, AC-24 (1979), pp. 913–921.

[27] J. S. MACNERNEY, *Investigation concerning positive definite continued fractions*, Duke Math. J., 26 (1959), pp. 663–677.

[28] E. P. MERKES AND W. T. SCOTT, *Periodic and reverse periodic continued fractions*, Michigan Math. J., 7 (1960), pp. 23–29.

[29] F. W. J. OLVER AND D. J. SOOKNE, *Note on backward recurrence algorithms*, Math. Comput., 26 (1972), pp. 941–947.

[30] O. PERRON, *Die Lehre von den Kettenbrüchen*, Second Edition, Tuebner, Leipzig, 1929; reprinted by Chelsea, New York, 1950.

[31] ———, *Die Lehre von den Kettenbrüchen*, Third Ed., Volume II, Tuebner, Stuttgart, 1957.

[32] A. PETERSON AND J. RIDENHOUR, *Oscillation of second order linear matrix difference equations*, J. Differential Equations, 89 (1991), pp. 69–88.

[33] S. PINCHERLE, *Delle funzioni ipergeometriche e di varie questioni ad esse attinenti*, Giornale di Mathematiche di Battaglini, 32, 1894, pp. 228–230.

[34] J. E. POTTER, *Matrix quadratic solutions*, SIAM J. Appl. Math., 14 (1966), pp. 496–501.

[35] W. T. REID, *Oscillation criteria for linear differential systems with complex coefficients*, Pacific J. Math., 6 (1956), pp. 147–169.

[36] ———, *Principal solutions of non-oscillatory, self-adjoint linear differential systems*, Pacific J. Math., 8 (1958), pp. 147–169.

[37] ———, *Riccati matrix differential equations and non-oscillation criteria for associated linear differential systems*, Pacific J. Math., 10 (1963), pp. 665–685

[38] F. RYDE, *Arithmetical Continued Fractions*, Lunds Universitets Årsskrift, N.F. Avd. 2, 22 (1926), pp. 1–182.

[39] G. SANSONE, *Equazioni Differenziali nel Campo Reale,* Parte Seconda, seconda ed., Nicola Zanichelli, Editore, Bologna, 1949.

[40] H. SCHWERDTFEGER, *Moebius transformations and continued fractions*, Bull. Amer. Math. Soc., 52 (1946), pp. 307–309.

[41] ———, *Geometry of Complex Numbers*, University of Toronto Press, Toronto, 1962.

[42] C. L. SIEGEL, *Symplectic Geometry*, Academic Press, New York, 1964.

[43] R. A. SMITH, *Matrix equation $XA + BX = C$*, SIAM J. Appl. Math., 16 (1968), pp. 198–201.

[44] D. R. VAUGHAN, *A nonrecursive algebraic solution for the discrete Riccati equation*, IEEE Trans. Automatic Control, 15 (1970), pp. 597–599.

[45] H. S. WALL, *Analytic Theory of Continued Fractions*, Van Nostrand, New York, 1948; reprinted by Chelsea, New York, 1973.

SIAM J. MATH. ANAL.
Vol. 24, No. 6, pp. 1622–1651, November 1993

011

MORE BOUNDS ON EIGENVALUE RATIOS FOR DIRICHLET LAPLACIANS IN N DIMENSIONS*

MARK S. ASHBAUGH[†] AND RAFAEL D. BENGURIA[‡]

Abstract. The authors investigate bounds for various combinations of the low eigenvalues of the Laplacian with Dirichlet boundary conditions on a bounded domain $\Omega \subset \mathbb{R}^n$. These investigations continue and expand upon earlier work of Payne, Pólya, Weinberger, Brands, Chiti, and the authors of this present paper. In particular, the authors generalize and extend to the n-dimensional setting various bounds of Payne, Pólya, Weinberger, Brands, and Chiti and examine their consequences and interrelationships in detail. This includes comparing the asymptotic forms of the various bounds as the dimension n becomes large. The authors also present various extensions and consequences of their recent proof of the Payne–Pólya–Weinberger conjecture, including the proof of a second conjecture of Payne, Pólya, and Weinberger under an added symmetry condition.

Key words. eigenvalues of Dirichlet Laplacians, universal eigenvalue inequalities, zeros of Bessel functions

AMS subject classifications. 35P15, 49Gxx, 35J05, 33A40

1. Introduction.

In this paper we go beyond our work in [6] to investigate such quantities as λ_3/λ_1 and $(\lambda_2 + \lambda_3)/\lambda_1$ for a bounded domain $\Omega \subset \mathbb{R}^n$. Here the quantity λ_i (for $i = 1, 2, \ldots$) denotes the ith eigenvalue of the Laplacian on Ω with zero Dirichlet boundary conditions imposed on $\partial\Omega$. In [6] (see also the announcement in [5]) we proved the Payne–Pólya–Weinberger conjecture in general dimension n, which says that

$$\lambda_2/\lambda_1\Big|_{\Omega} \leq \lambda_2/\lambda_1\Big|_{n\text{-dimensional ball}}, \tag{1.1}$$

with equality if and only if Ω is an n-dimensional ball. This bound had been conjectured by Payne, Pólya, and Weinberger (PPW) in [34] and [35], where they had proved that $\lambda_2/\lambda_1 \leq 3$ in the two-dimensional case (by contrast, (1.1) above gives approximately 2.5387 in this case). Thompson [47] made the natural extension of the results and conjecture of Payne, Pólya, and Weinberger to the n-dimensional setting. Between the papers of PPW and our own, several authors managed to lower the bound 3 in the PPW result for λ_2/λ_1 in two dimensions. These include Brands [12] who obtained 2.686 in 1964, de Vries [18] who obtained 2.658 in 1967, and Chiti [17] who obtained 2.586 in 1983. Only Chiti gave an n-dimensional version of his inequality, and even then he did not evaluate certain integrals of Bessel functions that occur in his bounds (except when $n = 2$). In this paper (§5) we reduce Chiti's n-dimensional result to its simplest form and compare his result with our own. We also extend

*Received by the editors October 13, 1992; accepted for publication (in revised form) January 5, 1993.

[†] Department of Mathematics, University of Missouri, Columbia, Missouri 65211 (mathmsa@mizzou1.missouri.edu). This author's work was partially supported by Fondo Nacional de Ciencia y Technologia (Chile) project number 0132-88 and by a Summer Research Fellowship provided by the Research Council of the University of Missouri-Columbia.

[‡] Facultad de Física, P. Universidad Católica de Chile, Avda. Vicuña Mackenna 4860, Casilla 306, Santiago 22, Chile (rbenguri@lascar.puc.cl). This author's work was supported in part by Fondo Nacional de Ciencia y Technologia (Chile) project numbers 0132-88 and 1238-90.

the basic result of Brands to the n-dimensional setting, recovering as a by-product a bound on λ_2/λ_1 which was given by Hile and Protter [23] (and which is a moderate improvement upon the general bound $\lambda_2/\lambda_1 \leq 1 + 4/n$ of PPW/Thompson).

In their original paper [35], Payne, Pólya, and Weinberger also formulated conjectures concerning λ_{m+1}/λ_m for $m = 2, 3, 4, \ldots$, and $(\lambda_2 + \lambda_3)/\lambda_1$ in two dimensions. These conjectures were that

(A) Among all bounded domains Ω and for all $m = 1, 2, 3, \ldots$, λ_{m+1}/λ_m is maximized by $\Omega =$ disk and $m = 1$ with equality only in this case, i.e.,

$$\lambda_{m+1}/\lambda_m \leq \lambda_2/\lambda_1\Big|_{\Omega=\text{disk}} \approx 2.5387; \tag{1.2}$$

(B) Among all bounded domains Ω, $(\lambda_2 + \lambda_3)/\lambda_1$ is maximized by $\Omega =$ disk, i.e.,

$$(\lambda_2 + \lambda_3)/\lambda_1 \leq (\lambda_2 + \lambda_3)/\lambda_1\Big|_{\Omega=\text{disk}} \approx 5.077. \tag{1.3}$$

These and related questions have been called attention to recently by Shubin in [3]. In our recent work [7], [8], we have extended our proof of the PPW conjecture to prove the $m = 2$ and $m = 3$ cases of (A) and even to prove that $\lambda_4/\lambda_2 < \lambda_2/\lambda_1$ (disk) ≈ 2.5387. However, the cases for $m > 3$ are still left open. In this paper, we concentrate on (B) and, while we do not prove it, we prove a special case of it for domains with 4-fold rotational symmetry. We also formulate n-dimensional generalizations of it, prove certain related but weaker inequalities, and examine the relations between our inequalities and the conjectured inequalities particularly in the limit as the dimension n goes to infinity.

The original papers of Payne, Pólya, and Weinberger have spawned a large collection of papers on "universal inequalites" between eigenvalues. A relatively complete set of references to this literature is contained in [6]. In the meantime we have become aware of a paper by Anghel [2] and of recent work of Harrell [20]. Particularly relevant to the present paper are the review articles by Protter [39], [41] (see also Payne's review article [32] for a more general overview as of 1967). Finally, we cannot leave our historical summary without at least mentioning the work on λ_3/λ_1 and $(\lambda_2 + \lambda_3)/\lambda_1$ in two dimensions which was begun by Payne, Pólya, and Weinberger and continued by Brands [12], Hile and Protter [23], and Marcellini [27]. Marcellini's result $\lambda_3/\lambda_1 \leq 7(15 + \sqrt{345})/60 \approx 3.9170$ is the best bound on λ_3/λ_1 to date (Brands found $\lambda_3/\lambda_1 \leq (7 + 2\sqrt{7})/3 \approx 4.097$ while Hile and Protter found $\lambda_3/\lambda_1 \leq 4.014$). Similarly, Marcellini's bound $(\lambda_2 + \lambda_3)/\lambda_1 \leq (15 + \sqrt{345})/6 \approx 5.5957$ is the best general bound on $(\lambda_2 + \lambda_3)/\lambda_1$ to date (Brands found $(\lambda_2 + \lambda_3)/\lambda_1 \leq 3 + \sqrt{7} \approx 5.646$ while Hile and Protter found $(\lambda_2 + \lambda_3)/\lambda_1 \leq 5.622$). These bounds improve upon the bound $(\lambda_2 + \lambda_3)/\lambda_1 \leq 6$ proved by Payne, Pólya, and Weinberger [35] (who also observed from this that $\lambda_3/\lambda_1 < 5$ but could easily have obtained $\lambda_3/\lambda_1 \leq 13/3 \approx 4.333$ by using $(\lambda_2 + \lambda_3)/\lambda_1 \leq 6$ in combination with another of their results, $\lambda_3 \leq \lambda_1 + 2\lambda_2$). The work of these authors was largely confined to the two-dimensional case though certain of the results of Brands and Payne, Pólya, and Weinberger do generalize fairly easily (see inequalities (6.2) and (6.10) in §6 below). In addition, many of the results of Hile and Protter are obtained for general dimension n. Work on λ_3/λ_1 is made more difficult by the fact that no reasonable guess exists as to the precise shape of the domain which maximizes it.

We also examine various consequences of our proof of the PPW conjecture for λ_2/λ_1. *In particular, we show that it leads to optimal inequalities for quantities* considered by Pólya [36] and to an improved upper bound in an inequality for the gap $\lambda_2 - \lambda_1$ due to Singer, Wong, Yau, and Yau [45]. These results are in our §§4 and 7, respectively.

2. Notational preliminaries. In this section we summarize our notation and recall from [6] some of the more useful formulas for our present considerations. We deal always with the Dirichlet eigenvalue problem for the Laplacian on a bounded domain $\Omega \subset \mathbb{R}^n$,

$$-\Delta u = \lambda u \quad \text{on } \Omega \tag{2.1}$$

with boundary condition

$$u = 0 \quad \text{on } \partial\Omega. \tag{2.2}$$

We list the eigenvalues (with multiplicity) as

$$0 < \lambda_1 < \lambda_2 \leq \lambda_3 \leq \cdots \to \infty \tag{2.3}$$

and a corresponding complete orthonormal set of eigenfunctions will be denoted by $\{u_i\}_{i=1}^{\infty}$. The Rayleigh–Ritz characterization of the eigenvalues will be an ever-present tool. One has

$$\lambda_i = \min_{\substack{\varphi \in H_0^1(\Omega)\setminus\{0\} \\ \varphi \perp W_{i-1}}} \frac{\int_\Omega |\nabla\varphi|^2 dx}{\int_\Omega \varphi^2 dx}, \tag{2.4}$$

where $W_i =$ span $\{u_1, \dots, u_i\}$ (with $W_0 = \{0\}$). Here dx represents the n-dimensional Lebesgue measure in $\mathbb{R}^n$. For our purposes no harm will be done by restricting consideration to real-valued functions; thus we shall dismiss all complex conjugations from our inner products. We shall be particularly concerned with trial functions φ of the form Pu_1. For these it is useful to formulate (2.4) as

$$\lambda_i - \lambda_1 \leq \min_{\substack{\varphi = Pu_1 \neq 0 \\ \varphi \in W_{i-1}^{\perp}}} \frac{\int_\Omega |\nabla P|^2 u_1^2 dx}{\int_\Omega P^2 u_1^2 dx}. \tag{2.5}$$

There is also a trace version of the Rayleigh–Ritz principle which we refer to as the extended Rayleigh–Ritz principle:

$$\sum_{k=1}^{m} \lambda_k \leq Tr\ A(\varphi_1, \dots, \varphi_m), \tag{2.6}$$

where $A(\varphi_1, \dots, \varphi_m)$ denotes the matrix whose (i, j)th entry is $(\varphi_i, (-\Delta)\varphi_j)$ and $\{\varphi_i\}_{i=1}^{m}$ is any orthonormal set of m vectors from the underlying Hilbert space.

We shall be particularly concerned with trial functions

$$P_i = g(r)x_i/r \quad \text{for } i = 1, 2, \dots, n, \tag{2.7}$$

where the x_i's are the standard Cartesian coordinates and $r = |x|$. A simple topological argument based on the Brouwer fixed point theorem guarantees that the origin can be chosen so that $P_i u_1 \perp u_1$ for all $i = 1, \dots, n$ if g is nonnegative. This argument is due originally to Weinberger [49] who used it in an analogous situation for the Neumann eigenvalue problem. A further topological argument based on the Borsuk–Ulam theorem will be developed below. It will allow us to rotate axes so that various higher orthogonality constraints are satisfied for at least some of the P_i's (in particular, it will guarantee $P_k u_1 \in W_k^\perp$ for $k = 1, 2, \dots, n$ if the axes are ordered appropriately). Often we take $g(r) = r$. In that case $P_i = x_i$ and the Rayleigh quotients in (2.5) become simply

$$\frac{\int_\Omega u_1^2 dx}{\int_\Omega x_i^2 u_1^2 dx}. \tag{2.8}$$

In this case the Brouwer fixed point theorem argument can be replaced simply by a center of mass argument. When we can sum the inequalities arising from all the P_i's (where $P_i = g(r)x_i/r$ again) it is useful to note that

$$\sum_{i=1}^n P_i^2 = g(r)^2 \tag{2.9}$$

and

$$\sum_{i=1}^n |\nabla P_i|^2 = g'(r)^2 + (n-1)\frac{g(r)^2}{r^2}. \tag{2.10}$$

In particular, when $P_i u_1 \in W_1^\perp$ for all $i = 1, \dots, n$ this gives the basic gap inequality

$$\lambda_2 - \lambda_1 \le \frac{\int_\Omega \left[g'(r)^2 + (n-1)g(r)^2/r^2\right] u_1^2 dx}{\int_\Omega g(r)^2 u_1^2 dx}. \tag{2.11}$$

We shall also need Bessel functions and their zeros. We use the standard notation of Abramowitz and Stegun [1]. The standard Bessel function of the first kind of order p will be denoted $J_p(x)$, and its kth positive zero will be denoted $j_{p,k}$. For an n-dimensional ball of radius R the low Dirichlet eigenvalues are $\lambda_1 = j_{n/2-1,1}^2/R^2$ and $\lambda_2 = \cdots = \lambda_{n+1} = j_{n/2,1}^2/R^2$. Corresponding (unnormalized but orthogonal) eigenfunctions are $u_1 = r^{1-n/2}J_{n/2-1}(j_{n/2-1,1}r/R)$ and $u_{i+1} = r^{1-n/2}J_{n/2}(j_{n/2,1}r/R)x_i/r$ for $i = 1, 2, \dots, n$. In particular, the bound in (1.1) is $(j_{n/2,1}/j_{n/2-1,1})^2$.

Rearrangement results will be used as needed. For most of the technical details for these and related matters we refer to [6] and the references therein. A sampling of these references includes [11], [19], and [38]; in particular, [38] is largely devoted to applying rearrangement results to the eigenvalue problems of mathematical physics (see also Bandle's book [11] for much useful information on rearrangement inequalities and eigenvalue problems). Chavel's book [14] also contains related material, particularly with reference to generalizing various results to Riemannian manifolds (usually with constraints of some sort on their curvature).

3. A second conjecture of Payne, Pólya, and Weinberger: Proof for domains with symmetry of order 4. A second conjecture of Payne, Pólya, and Weinberger, contained in [35], is that for the Dirichlet Laplacian on a bounded domain $\Omega \subset \mathbb{R}^2$,

$$(\lambda_2 + \lambda_3)/\lambda_1 \le 2j_{1,1}^2/j_{0,1}^2 \tag{3.1}$$

with equality if and only if Ω is a disk. In [35], the bound 6 was established for $(\lambda_2 + \lambda_3)/\lambda_1$, whereas the conjectured bound given above has a numerical value of approximately 5.077. Various authors have brought the constant here down as follows: Brands [12] 5.646, Hile and Protter [23] 5.622, and Marcellini [27] 5.596. Here we show how our method of proof for the first PPW conjecture can be extended to prove (3.1) in the case that Ω has a center of symmetry of order 4. That is, we establish the following.

THEOREM 3.1. *Let Ω be a bounded domain in $\mathbb{R}^2$ with 4-fold rotational symmetry about a point. Then inequality* (3.1) *holds between the first three eigenvalues of the Laplacian on this domain with Dirichlet boundary conditions. Equality occurs if and only if Ω is a disk.*

Proof. The proof proceeds much like our proof [5], [6] of the Payne–Pólya–Weinberger conjecture for λ_2/λ_1. We start from the extended Rayleigh–Ritz inequality

$$\lambda_2 + \lambda_3 - 2\lambda_1 \leq \frac{\int_\Omega |\nabla P_1|^2 u_1^2 dx}{\int_\Omega P_1^2 u_1^2 dx} + \frac{\int_\Omega |\nabla P_2| u_1^2 dx}{\int_\Omega P_2^2 u_1^2 dx}, \tag{3.2}$$

where now P_1 and P_2 must be nontrivial functions such that $P_1 u_1$ and $P_2 u_1$ are both orthogonal to u_1 and to each other. Taking P_1 and P_2 as in the prior proof (i.e., of the form given by (2.7) with $g(r)$ chosen to be a particular ratio of Bessel functions), orthogonality to u_1 is assured by the symmetry of Ω (which is shared by u_1). To establish orthogonality to each other note that

$$\int_\Omega P_1 P_2 u_1^2 dx = \int_\Omega g^2(r) \left[\frac{1}{2}\sin 2\theta\right] u_1^2 dx = 0 \tag{3.3}$$

since rotation by 90° leaves $g(r)$ and u_1 alone, whereas $\sin 2\theta$ is changed to its negative. A second (crucial) fact deriving from the 4-fold symmetry of Ω is

$$\int_\Omega P_1^2 u_1^2 dx = \int_\Omega P_2^2 u_1^2 dx \tag{3.4}$$

(the numerators in (3.2) above are equal also). This follows by considering the behavior of $\int_\Omega (P_1^2 - P_2^2) u_1^2 dx = \int_\Omega g^2(r) \cos 2\theta u_1^2 dx$ under rotation by 90°. This allows us to replace the right-hand side in (3.2) by

$$\frac{2\int_\Omega \left[g'(r)^2 + \frac{1}{r^2} g(r)^2\right] u_1^2 dx}{\int_\Omega g(r)^2 u_1^2 dx}$$

and then proceed through to the conclusion of the proof exactly as in [5] and [6], except now we have the extra factor of 2 on the right-hand side. □

Remark. One could also have proceeded by arguing from the Rayleigh–Ritz inequality in a slightly different fashion. We can obtain the two inequalities

$$\lambda_2 - \lambda_1 \leq \frac{\int_\Omega |\nabla P_1|^2 u_1^2 dx}{\int_\Omega P_1^2 u_1^2 dx} \tag{3.5}$$

and

$$\lambda_3 - \lambda_1 \leq \frac{\int_\Omega |\nabla P_2|^2 u_1^2 dx}{\int_\Omega P_2^2 u_1^2 dx} \tag{3.6}$$

if we choose P_1 and P_2 as before and also use rotational freedom to fix our coordinate axes so that P_2u_1 is orthogonal to u_2. This is possible because under 180° rotation u_2 can be taken to be either even or odd. In the even case orthogonality is automatic (i.e., no choice of axes is required). In the odd case we observe that if $\int_\Omega P_2u_1u_2dx \neq 0$, then after a 180° rotation of axes this quantity will change signs and hence, by continuity, for some intermediate rotation orthogonality will obtain. Having established (3.5) and (3.6) we can sum them and proceed as before. The 4-fold rotational symmetry is needed to establish (3.4); thus, in a certain sense the consequence (3.4) of the 4-fold rotational symmetry is more crucial to establishing (3.1) than the orthogonality condition (3.3).

In fact, from the alternative approach presented in the preceding remark we can actually obtain the separate inequalities given by the following.

THEOREM 3.2. *With hypotheses as in Theorem* 3.1,

$$\lambda_i/\lambda_1 \leq j_{1,1}^2/j_{0,1}^2 \quad \textit{for } i = 2,3. \tag{3.7}$$

For the proof one simply observes that

$$\int_\Omega |\nabla P_i|^2 u_1^2 dx = \frac{1}{2}\int_\Omega \left[|\nabla P_1|^2 + |\nabla P_2|^2\right] u_1^2 dx \quad \text{for } i = 1,2 \tag{3.8}$$

and

$$\int_\Omega P_i^2 u_i^2 dx = \frac{1}{2}\int_\Omega \left[P_1^2 + P_2^2\right] u_1^2 dx \quad \text{for } i = 1,2. \tag{3.9}$$

The results above have a connection to a celebrated conjecture concerning the nodal line of the second eigenfunction for the Dirichlet problem in two dimensions. The conjecture states that the nodal line of u_2 must cross Ω, i.e., that u_2 cannot have a closed nodal line not touching $\partial\Omega$ (for statements and discussion see [32], [40], [50], and [51]). Throughout our discussion of the nodal line conjecture and its implications for our problem we shall assume that Ω has a C^1 boundary, i.e., that each component of the boundary is given by a simple, closed curve having a continuously turning unit tangent vector. Though partial results have been obtained for this problem they have all required additional assumptions of symmetry and/or convexity [24], [25], [29], [33], [42]. Examples of these results are that u_2 cannot have a closed interior nodal line if Ω is convex and symmetric with respect to a line (Payne [33]) or if Ω is convex and has a discrete rotational symmetry (Lin [25]). Other results relating to this conjecture are contained in the papers of Shen [42] and Jerison [24]. Very recently, Melas [29] has obtained the strongest result to date, requiring only convexity. His work is based on the prior work of Lin [25] and Payne [33].

Based on Lin's result (now a special case of Melas' result [29]) and an observation of Weinberger (made in connection with our work on Neumann eigenvalues in [9]) we can prove

THEOREM 3.3. *Let $\Omega \subset \mathbb{R}^2$ be a bounded convex domain with a smooth boundary and having k-fold rotational symmetry with $k \geq 3$. Then $\lambda_2 = \lambda_3$ and hence inequalities* (3.1) *and* (3.7) *also hold.*

Proof. Aside from the condition $k \geq 3$, the hypotheses here are those of Lin [25]. Therefore, u_2 must have a nodal line which touches the boundary at exactly two points. This shows that u_2 cannot be invariant under rotation by $2\pi/k$ radians and hence that λ_2 is at least doubly degenerate (since u_2 rotated by $2\pi/k$ radians must be an independent eigenfunction for λ_2). □

We note that Theorem 3.3 renders Theorems 3.1 and 3.2 somewhat trivial for the cases that it covers (and also that it goes beyond these theorems in the respect that it allows k's not divisible by 4). On the other hand, Theorems 3.1 and 3.2 are more general in that Ω is allowed to be nonconvex and even nonsimply connected. In addition, one can go beyond what we have done above in the following way, which might actually be of use in establishing the nodal line conjecture for all domains with 4-fold rotational symmetry.

THEOREM 3.4. *Let $\Omega \subset \mathbb{R}^2$ have 4-fold rotational symmetry and suppose that u_2 is even, i.e., $u_2(-x) = u_2(x)$ for all $x \in \Omega$. Then*

$$\lambda_i/\lambda_1 \leq j_{1,1}^2/j_{0,1}^2 \quad \textit{for } i = 3, 4 \tag{3.10}$$

and hence also $(\lambda_3 + \lambda_4)/\lambda_1 \leq 2j_{1,1}^2/j_{0,1}^2$.

Note. This theorem makes no assumptions concerning the connectivity of Ω or the smoothness of its boundary.

Proof. One simply observes that if u_2 is even, then both P_1u_1 and P_2u_1 will be orthogonal to it (see, in particular, the arguments given in the remark following Theorem 3.1). Then all the estimates given above are seen to apply to P_1u_1 and P_2u_1 as trial functions for u_3 and u_4 (the latter by virtue of a rotation, if necessary). In particular, (3.5) and (3.6) can be applied, with λ_3 and λ_4 replacing λ_2 and λ_3 in their left-hand sides, and together with (3.8) and (3.9) these yield (3.10) of our theorem. □

Therefore, if there were a domain with 4-fold symmetry for which the nodal line conjecture were false we would have $\lambda_4/\lambda_1 \leq j_{1,1}^2/j_{0,1}^2$ (since in this context it is clear that if u_2 has a closed interior nodal line, then u_2 can be taken to be even). This may perhaps be a useful result to have in working to show that u_2 cannot have a closed interior nodal line (and hence must be odd). In fact, it seems reasonable to believe that u_2 for a nonconvex domain would be more likely to have a crossing nodal line than u_2 for a convex domain. This certainly suggests that the nodal line conjecture should hold for all domains with 4-fold rotational symmetry but we repeat that this has not yet been proved. It seems that we are in the somewhat paradoxical situation that the harder (convex) case is proved while the nonconvex case remains unproved. If the nodal line conjecture turns out to be true, then for simply connected domains having 4-fold rotational symmetry and smooth boundaries our Theorem 3.4 above would hold vacuously.

The n-dimensional analog of Theorem 3.1 concerns the quantity

$$(\lambda_2 + \lambda_3 + \cdots + \lambda_{n+1})/\lambda_1$$

(but see also our comments near the end of §5) and deals with bounded domains $\Omega \subset \mathbb{R}^n$ for which there is a choice of Cartesian coordinates such that Ω is invariant with respect to 90° rotation in each of the $\binom{n}{2}$ coordinate planes. We have the following.

THEOREM 3.5. *Let Ω be a bounded domain in $\mathbb{R}^n$ and suppose Ω is invariant with respect to 90° rotations in the coordinate planes spanned by each pair of (Cartesian) coordinate axes. Then*

$$(\lambda_2 + \lambda_3 + \cdots + \lambda_{n+1})/\lambda_1 \leq n\left(j_{n/2,1}/j_{n/2-1,1}\right)^2 \tag{3.11}$$

holds and, furthermore, equality obtains if and only if Ω is an n-dimensional ball.

Proof. The proof here is essentially identical to our previous one. One takes $P_i = g(r)x_i/r$, where x_i is a Cartesian coordinate. Orthogonality of each of the functions $P_i u_1$ $(1 \le i \le n)$ to u_1 is assured by the symmetry of Ω. Pairwise orthogonality of the $P_i u_1$'s $(1 \le i \le n)$ follows as for (3.3) above by making a 90° rotation in the $x_i x_j$-plane, where i and j are the indices of the relevant P's. Finally,

$$\int_\Omega P_1^2 u_1^2 dx = \int_\Omega P_2^2 u_1^2 dx = \cdots = \int_\Omega P_n^2 u_1^2 dx,$$

the analog of (3.4) above, also follows from the 4-fold rotational symmetry of Ω in each coordinate plane.

By the extended Rayleigh–Ritz inequality one then has

$$\begin{aligned}(\lambda_2 + \lambda_3 + \cdots + \lambda_{n+1}) - n\lambda_1 &\le \sum_{i=1}^{n} \frac{\int_\Omega |\nabla P_i|^2 u_1^2 dx}{\int_\Omega P_i^2 u_1^2 dx} \\ &= \frac{n \int_\Omega \sum_{i=1}^{n} |\nabla P_i|^2 u_i^2 dx}{\int_\Omega \sum_{i=1}^{n} P_i^2 u_1^2 dx} \\ &= n\frac{\int_\Omega \left[g'(r)^2 + (n-1)g(r)^2/r^2\right] u_1^2 dx}{\int_\Omega g(r)^2 u_1^2 dx},\end{aligned} \tag{3.12}$$

and, as in our previous papers [5], [6], this ultimately yields

$$(\lambda_2 + \lambda_3 + \cdots + \lambda_{n+1}) - n\lambda_1 \le n\lambda_1 \left[\left(j_{n/2,1}/j_{n/2-1,1}\right)^2 - 1\right],$$

which is equivalent to (3.11). Characterization of the cases of equality also follows our prior work. □

For weaker inequalities than (3.11) but which hold for general domains (i.e., no symmetry hypothesis required) see §5. There we discuss the relations between these weaker inequalities, the conjectured inequality (3.11) for arbitrary domains $\Omega \subset \mathbb{R}^n$, and an intermediate inequality (also conjectural except under our 4-fold symmetry hypothesis, at this point). Also, in §6 we obtain (nonoptimal) bounds on $(\lambda_2 + \cdots + \lambda_{n+1})/\lambda_1$ which are the n-dimensional analogs of the two-dimensional results $(\lambda_2 + \lambda_3)/\lambda_1 \le 6$ of PPW and $(\lambda_2 + \lambda_3)/\lambda_1 \le 5 + \lambda_1/\lambda_2$ of Brands and investigate some of their consequences.

The results of this section have obvious extensions to the case of Schrödinger operators and quite general second-order elliptic equations (as discussed in §4 of [6]) under suitable symmetry conditions, i.e., if all the coefficient functions share the 4-fold symmetries of Ω.

4. Other isoperimetric results implied by our result.

In this section we give two isoperimetric inequalities which follow from our main result and previously known results. These have been discussed in the previous literature as conjectures. In particular, we recommend the article of Hersch [21] and also his comments and updates to Pólya's papers in [22] (see the comments to papers 202 and 203, specifically).

In [36], Pólya asked for the least upper bound for the quantity

$$\lambda_2 \dot{r}^2$$

and for the shape of domain for which this bound is attained or approximated (see question 5(c) on p. 336 of [36]). Here $\dot{r}$ denotes the maximum inner conformal radius of the domain $\Omega \subset \mathbb{R}^2$, which is here supposed to be a simply connected domain (we use our notation for λ_2 here rather than Pólya's). By virtue of our bound on λ_2/λ_1 and the Pólya–Szegö bound [38, pp. 97–98]

$$\lambda_1 \dot{r}^2 \leq j_{0,1}^2 \approx 5.7832 \tag{4.1}$$

(which is isoperimetric, with equality only for disks), the isoperimetric inequality

$$\lambda_2 \dot{r}^2 \leq j_{1,1}^2 \approx 14.6820 \tag{4.2}$$

follows directly (with equality if and only if Ω is a disk), thus answering Pólya's question in its entirety. Pólya himself only explicitly gave the bound $j_{0,2}^2 \approx 30.4713$. With the PPW result $\lambda_2/\lambda_1 \leq 3$ this value could be reduced to $3j_{0,1}^2 \approx 17.3496$, and obviously the improvements in the bound for λ_2/λ_1 as found by Brands, de Vries, and Chiti all lead to further reductions. One can even do slightly better each time if instead of using the bound for λ_2/λ_1 with the Pólya–Szegö bound (4.1) above one uses it with the isoperimetric bound (equality if and only if Ω is a disk)

$$\left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2}\right)\frac{1}{\dot{r}^2} \geq \frac{1}{j_{0,1}^2} + \frac{1}{j_{1,1}^2} \tag{4.3}$$

of Pólya and Schiffer [37], a fact which appears to have first been utilized by Hersch [21]. Hersch, using $\lambda_2/\lambda_1 \leq 3$, then derived the bound $\lambda_2 \dot{r}^2 \leq 16.5957$. Using Chiti's bound, $\lambda_2/\lambda_1 \leq 2.585965$, the constant here would improve to 14.8779, which is the best value derivable prior to our proof of the PPW conjecture. Of course, our isoperimetric bound (4.2) follows by combining our bound $\lambda_2/\lambda_1 \leq j_{1,1}^2/j_{0,1}^2$ with either of the bounds (4.1) or (4.3).

A second result of this nature which follows from our proof of Payne, Pólya, and Weinberger's second conjecture for the case of domains possessing symmetry of order 4 (see §3 above) is the bound

$$(\lambda_2 + \lambda_3)\dot{r}^2 \leq 2j_{1,1}^2 \approx 29.3639. \tag{4.4}$$

This holds for any simply connected domain in the plane having symmetry of order 4. The best general result of this type (i.e., without symmetry assumptions) known at present is

$$(\lambda_2 + \lambda_3)\dot{r}^2 \leq 32.36095^+, \tag{4.5}$$

which follows from Marcellini's bound $(\lambda_2 + \lambda_3)/\lambda_1 \leq (15 + \sqrt{345})/6 \approx 5.5957$. Obviously, (4.5) is not expected to be isoperimetric.

In addition, for convex domains in $\mathbb{R}^2$ having smooth boundary and k-fold rotational symmetry with $k \geq 3$ one has, by Theorem 3.3 above, $\lambda_2 = \lambda_3$ so that (4.4) and also the stronger inequalities

$$\lambda_i \dot{r}^2 \leq j_{1,1}^2 \approx 14.6820 \quad \text{for } i = 2, 3 \tag{4.6}$$

hold in this case. From Theorem 3.2 one sees that the inequalities (4.6) also hold for any simply connected domain having 4-fold rotational symmetry, whether or not it is convex with a smooth boundary.

Further inequalities with $\dot{r}$ replaced by d, where d is defined as the radius of the largest disk contained in Ω, follow from the inequality $d \leq \dot{r} \leq 4d$ of Koebe. Another useful inequality for such considerations is $\dot{r}^2 \leq A/\pi \leq \overline{r}^2$ [38, p.8], where $\overline{r}$ is the outer conformal radius of Ω and A is its area.

Note that the entire discussion in this section is confined to the two-dimensional case.

5. Large n asymptotics and Chiti's bound. In this section we give an asymptotic expansion for our bound $\left(j_{n/2,1}/j_{n/2-1,1}\right)^2$ for large dimension n and compare it with the bounds of PPW/Thompson [34], [35], [47] and Chiti [17]. We also prove some (nonisoperimetric) bounds which are extensions of Chiti's bound and examine various ramifications with respect to certain further conjectures of Payne, Pólya, and Weinberger (see §3 also).

Using the asymptotic development for $j_{p,1}$ for large p as given by Abramowitz and Stegun [1, p. 371, eq. 9.5.14] it is easy to work out that

$$\left(\frac{j_{n/2,1}}{j_{n/2-1,1}}\right)^2 = 1 + \frac{4}{n} - \frac{4}{3}(1.8557571)\frac{2^{5/3}}{n^{5/3}} + \frac{12}{n^2} + O\left(n^{-7/3}\right). \tag{5.1}$$

(The value 1.8557571 appearing here represents $|a_1|/2^{1/3}$, where a_1 is the first negative zero of the Airy function $Ai(x)$; see Olver [31, p. 408, Ex. 6.4] or Abramowitz and Stegun [1, p. 371, eq. 9.5.22 and p. 368, eq. 9.3.39].) Incidentally, this shows that the PPW/Thompson bound $1 + 4/n$ is the best possible of the form $1 + c/n$: that is, if one seeks a constant c (independent of n) such that $\lambda_2/\lambda_1 \leq 1 + c/n$ holds for all dimensions n, then the value $c = 4$ of PPW/Thompson is the best possible.

In arbitrary dimension n the bound of Chiti is

$$\lambda_2 \leq \left(1 + \frac{n}{2}\frac{J^2_{n/2}(j_{n/2-1,1})}{j^2_{n/2-1,1}\int_0^1 r^3 J^2_{n/2-1}(j_{n/2-1,1}r)\,dr}\right)\lambda_1 \tag{5.2}$$

as given in [17] and reproduced in [41]. The integral occurring here is not too difficult to work out (see Appendix A), and one finds that (5.2) takes the much simpler form

$$\lambda_2/\lambda_1 \leq 1 + \frac{6n}{2j^2_{n/2-1,1} + n(n-4)}. \tag{5.3}$$

Again, the asymptotics of the right-hand side of this inequality may be worked out, yielding

$$1 + \frac{6n}{2j^2_{n/2-1,1} + n(n-4)} = 1 + \frac{4}{n} - \frac{4}{3}(1.8557571)\frac{2^{5/3}}{n^{5/3}} + \frac{16}{n^2} + O(n^{-7/3}). \tag{5.4}$$

A comparison with the best bound as given by (5.1) shows that Chiti's bound gets the next term beyond $4/n$ in the asymptotic expansion "right."

It is clear from the asymptotic formulas that

$$\left(\frac{j_{n/2,1}}{j_{n/2-1,1}}\right)^2 < 1 + \frac{6n}{2j^2_{n/2-1,1} + n(n-4)} < 1 + \frac{4}{n} \tag{5.5}$$

for large dimension n. Indeed, the first of these inequalities holds for all n as follows from Chiti's bound. The second would hold for all n if the inequality

$$j_{p,1} > \sqrt{(p+1)(p+5)} \tag{5.6}$$

were to hold for $p = -1/2, 0, 1/2, 1, 3/2, \ldots$. This inequality has recently been shown to hold by Lorch [26] for all $p > -1$. Watson's book gives only

$$j_{p,1} > \sqrt{p(p+2)}$$

[48, p. 486, ineq. (5)]. Lorch in fact shows that $j_{p,1}^2 - (p+1)(p+5)$ increases from zero to infinity as p runs from -1 to infinity. Lorch also has shown that the bound $(n+3+\sqrt{n^2+10n+9})/2n$ due to Brands [12] (for $n = 2$) and Hile and Protter [23] (for general dimension n) is intermediate between Chiti's bound and the bound $1+4/n$ of Payne, Pólya, and Weinberger (cf. (5.5)). We discuss the Brands/Hile–Protter bound and Lorch's result concerning it in more detail in §6 below.

Finally, we observe that the arguments of Chiti in [16] and [17] as used in this paper allow one to prove the bound

$$\frac{1}{\lambda_2-\lambda_1}+\frac{1}{\lambda_3-\lambda_1} \geq \frac{j_{0,1}^2-2}{3\lambda_1}. \tag{5.7}$$

This is an improvement upon the two-dimensional case of (5.2) and also upon Chiti's bound

$$\frac{1}{\lambda_2-\lambda_1}+\frac{1}{\lambda_3-\lambda_1} \geq \frac{1}{\lambda_1} \tag{5.8}$$

which follows directly from [16] and rearrangement inequalities (since $\lambda_1 \geq \lambda_1^\star$ where $\lambda_1^\star$ is defined as the Rayleigh quotient for $-\Delta$ on Ω^* with $u_1^\star$ as trial function; here Ω^* denotes the disk of the same volume as Ω and $u_1^\star$ denotes the "spherical decreasing rearrangement" of u_1; see [5] and [6] for more details and references) because $\left(j_{0,1}^2-2\right)/3 \approx 1.2611 > 1$. The lower bounds of Chiti [16], $1/\lambda_1^\star$ and $\int_\Omega u_1^2 dx/2\pi u_M^2$, where $u_M^2 = (\text{max of } u_1^2)$, are not directly comparable to the bound in (5.7) above, though in any event they are less accessible as they require knowledge of u_1 or $u_1^\star$.

If the second conjecture (see §3 above),

$$(\lambda_2+\lambda_3)/\lambda_1 \leq 2j_{1,1}^2/j_{0,1}^2, \tag{5.9}$$

of Payne–Pólya–Weinberger were to hold, then (5.7) could be improved to

$$\frac{1}{\lambda_2-\lambda_1}+\frac{1}{\lambda_3-\lambda_1} \geq \frac{2j_{0,1}^2}{\left(j_{1,1}^2-j_{0,1}^2\right)\lambda_1} = \frac{2}{\left[\left(j_{1,1}/j_{0,1}\right)^2-1\right]\lambda_1}, \tag{5.10}$$

where now the constant on the right-hand side has the value $2/\left[\left(j_{1,1}/j_{0,1}\right)^2-1\right] \approx 1.2998$. The inequalities (5.9) and (5.10), if true, would be isoperimetric with equality only for the disk. In any event, (5.9) would imply (5.10) but not conversely so (5.10) is of intermediate strength between the known result (5.7) and the second inequality (5.9) conjectured by Payne, Pólya, and Weinberger. Note that either of the inequalities (5.9) or (5.10) would imply our result (1.1), i.e., the (first) PPW conjecture. Similarly, the bound (5.7) implies the two-dimensional bound that Chiti gave explicitly in [17] for λ_2/λ_1.

We now show how we arrive at (5.7) and then go on to its analog in higher dimensions. As in [12], [16], and [27] we observe that we can choose coordinate axes so that

$$\int_\Omega x u_1^2 dx = \int_\Omega y u_1^2 dx = \int_\Omega x u_1 u_2 dx = 0 \tag{5.11}$$

by suitable translation and rotation of our axes. It then follows by the Rayleigh–Ritz inequality with xu_1 and yu_1 as trial functions for u_3 and u_2 that

$$\lambda_2 - \lambda_1 \leq \frac{\int_\Omega u_1^2 dx}{\int_\Omega y^2 u_1^2 dx}$$

and

$$\lambda_3 - \lambda_1 \leq \frac{\int_\Omega u_1^2 dx}{\int_\Omega x^2 u_1^2 dx}.$$

From these,

$$\frac{1}{\lambda_2 - \lambda_1} + \frac{1}{\lambda_3 - \lambda_1} \geq \frac{\int_\Omega (x^2 + y^2) u_1^2 dx}{\int_\Omega u_1^2 dx} = \frac{\int_\Omega r^2 u_1^2 dx}{\int_\Omega u_1^2 dx} \tag{5.12}$$

follows immediately and then rearrangement inequalities and Chiti's comparison result (see [5] and [6], including Appendix A of [6], for more on this) do the rest to yield (5.7).

In higher dimensions the analog of (5.7) is

$$\frac{1}{\lambda_2 - \lambda_1} + \frac{1}{\lambda_3 - \lambda_1} + \cdots + \frac{1}{\lambda_{n+1} - \lambda_1} \geq \frac{2j_{n/2-1,1}^2 + n(n-4)}{6\lambda_1}, \tag{5.13}$$

where n denotes the dimension. This inequality follows in exact analogy to the two-dimensional case once one shows that by a proper choice of origin and axes one can arrange for the orthogonality conditions

$$\int_\Omega x_i u_1 u_j dx = 0 \quad \text{for } i = 1, 2, \ldots, n \quad \text{and} \quad j = 1, \ldots, n + 1 - i \tag{5.14}$$

to be simultaneously satisfied. That this is possible follows in stages. First, that

$$\int_\Omega x_i u_1^2 dx = 0 \quad \text{for } i = 1, \ldots, n$$

can be satisfied by an appropriate choice of origin follows from the Brouwer fixed point theorem as alluded to in §2; or, more simply, one can view this as choosing the origin to lie at the center of mass of a mass distribution in Ω having mass density u_1^2. Next, one considers

$$\int_\Omega x_1 u_1 u_j dx \qquad \text{for } j = 2, \ldots, n$$

as a function on S^{n-1}, the unit sphere in $\mathbb{R}^n$, where the variable $\sigma \in S^{n-1}$ represents the direction of the positive x_1 axis. This is a mapping f_1 from S^{n-1} to $\mathbb{R}^{n-1}$ (the index j gives the component) which is antipode preserving, i.e., $f_1(-\sigma) = -f_1(\sigma)$. It now follows from the Borsuk–Ulam theorem (see [46, p. 266], [28, p. 170], or [30, p. 361]) that there is a $\sigma_1 \in S^{n-1}$ at which f_1 vanishes (f_1 also vanishes at $-\sigma_1$). We can, therefore, fix the x_1 axis in the direction of σ_1 to guarantee that $\int_\Omega x_1 u_1 u_j dx = 0$ for $j = 2, \ldots, n$. Having achieved this we now repeat the process by considering

$$\int_\Omega x_2 u_1 u_j dx \quad \text{for } j = 2, \ldots, n-1$$

as a function on the sphere S^{n-2}, which is the equator of S^{n-1} consisting of those vectors in S^{n-1} that are perpendicular to the (now fixed) direction of the x_1 axis, i.e., to σ_1. Again, this gives a mapping $f_2 : \; S^{n-2} \to \mathbb{R}^{n-2}$ which is antipode-preserving, and the Borsuk–Ulam theorem tells us that there is a $\sigma_2 \in S^{n-2}$ at which $f_2(\sigma_2) = 0$. If we take the x_2 axis to lie in this direction, then $\int_\Omega x_2 u_1 u_j dx = 0$ for $j = 2, \ldots, n-1$. Obviously, this process may be continued until we have fixed the directions of all the coordinate axes such that all the orthogonality conditions (5.14) hold. Using the Rayleigh–Ritz inequality with $x_i u_1$ as a trial function for u_{n+2-i} one finds easily that

$$\lambda_i - \lambda_1 \le \frac{\int_\Omega u_1^2 dx}{\int_\Omega x_{n+2-i}^2 u_1^2 dx} \quad \text{for } i = 2, 3, \ldots, n+1, \tag{5.15}$$

and then (5.13) follows directly as in the two-dimensional case.

If one computes the large n asymptotic expansion of the bound (5.13) and compares it to that of the conjectured bound

$$\frac{1}{\lambda_2 - \lambda_1} + \frac{1}{\lambda_2 - \lambda_1} + \cdots + \frac{1}{\lambda_{n+1} - \lambda_n} \ge \frac{n}{\left[\left(j_{n/2,1}/j_{n/2-1,1}\right)^2 - 1\right] \lambda_1} \tag{5.16}$$

(the extension to n dimensions of (5.10); all comments above following (5.10) apply to (5.16) as well if generalized to the n-dimensional setting; see also §3 on the second Payne–Pólya–Weinberger conjecture), one finds

$$\left[2j_{n/2-1,1}^2 + n(n-4)\right] \Big/ 6 \sim \frac{n^2}{4} \left[1 + \frac{2}{3}(1.8557571)\frac{2^{2/3}}{n^{2/3}} - \frac{4}{n} + O(n^{-4/3})\right], \tag{5.17}$$

whereas

$$\frac{n}{\left[\left(j_{n/2,1}/j_{n/2-1,1}\right)^2 - 1\right]} \sim \frac{n^2}{4} \left[1 + \frac{2}{3}(1.8557571)\frac{2^{2/3}}{n^{2/3}} - \frac{2}{n} + O(n^{-4/3})\right]. \tag{5.18}$$

From this it seems that the Chiti bound has a relative error of only $O(1/n)$ for large n (assuming that (5.16) is indeed true).

Of course, any of the bounds (5.13), (5.16), or

$$\left(\lambda_2 + \lambda_3 + \cdots + \lambda_{n+1}\right) / \lambda_1 \le n \left(j_{n/2,1}/j_{n/2-1,1}\right)^2 \tag{5.19}$$

(the last two of which are only conjectures, in general, at this point) implies a corresponding bound for the eigenvalues $\lambda_2, \ldots, \lambda_k$ for any integer k, $2 \le k \le n+1$. In particular, (5.13) leads to

$$\frac{1}{\lambda_2 - \lambda_1} + \frac{1}{\lambda_3 - \lambda_1} + \cdots + \frac{1}{\lambda_k - \lambda_1} \ge \frac{k-1}{n} \frac{2j_{n/2-1,1}^2 + n(n-4)}{6\lambda_1}, \tag{5.20}$$

(5.16) would lead to

$$\frac{1}{\lambda_2 - \lambda_1} + \frac{1}{\lambda_3 - \lambda_1} + \cdots + \frac{1}{\lambda_k - \lambda_1} \ge \frac{k-1}{\left[\left(j_{n/2,1}/j_{n/2-1,1}\right)^2 - 1\right] \lambda_1}, \tag{5.21}$$

and (5.19) would lead to

$$(5.22)\qquad (\lambda_2+\cdots+\lambda_k)/\lambda_1 \le (k-1)\left[j_{n/2,1}/j_{j/2-1,1}\right]^2$$

for $2 \le k \le n+1$. These follow by just using the obvious facts that

$$\frac{1}{\lambda_k-\lambda_1} \le \frac{1}{k-2}\left(\frac{1}{\lambda_2-\lambda_1}+\frac{1}{\lambda_3-\lambda_1}+\cdots+\frac{1}{\lambda_{k-1}-\lambda_1}\right)$$

and

$$\lambda_k \ge (\lambda_2+\lambda_3+\cdots+\lambda_{k-1})/(k-2),$$

and by working inductively. For $k=2$ the inequalities (5.20), (5.21), and (5.22) reduce to the known inequalities for λ_2/λ_1. The second conjecture of Payne, Pólya, and Weinberger concerning $(\lambda_2+\lambda_3)/\lambda_1$ in two dimensions could be considered to generalize to either (5.19) or the $k=3$ case of (5.22) in n dimensions; we prefer to think of (5.19) as the proper generalization since it seems the more natural of the two (nevertheless, in our next section we work to bound $(\lambda_2+\lambda_3)/\lambda_1$ in all dimensions n).

Finally, we remark that every bound developed above holds unchanged for the Schrödinger operator $H=-\Delta+V(x)$ acting on a bounded domain $\Omega\subset\mathbb{R}^n$ with Dirichlet boundary conditions and with a positive potential V. With minor changes (see §4 of [6]) these bounds also extend to rather general elliptic eigenvalue problems. There are also many Schrödinger problems on unbounded domains to which our bounds apply (see [4] for a discussion in one dimension which applies without change here; see also Simon [43], [44] and Ashbaugh–Exner [10] for some more exotic problems where this extension has interesting applications).

6. Bounds on $(\lambda_2+\lambda_3+\cdots+\lambda_{n+1})/\lambda_1$ and their implications for λ_3/λ_1 and $(\lambda_2+\lambda_3)/\lambda_1$ in n dimensions.

In this section we return to consideration of bounds for $(\lambda_2+\lambda_3)/\lambda_1$ and, most especially, to their extension to n dimensions. We shall begin by looking at bounds on $(\lambda_2+\lambda_3+\cdots+\lambda_{n+1})/\lambda_1$ in n dimensions and then turn to their implications for λ_3/λ_1 and $(\lambda_2+\lambda_3)/\lambda_1$. We again examine asymptotics in n as n goes to infinity to help judge the effectiveness of our bounds.

We begin by remarking that the bound

$$(6.1)\qquad (\lambda_2+\lambda_3)/\lambda_1 \le 6$$

of Payne, Pólya, and Weinberger [35] readily extends to n dimensions as

$$(6.2)\qquad (\lambda_2+\lambda_3+\cdots+\lambda_{n+1})/\lambda_1 \le n+4 = n(1+4/n).$$

This is easily arrived at by using trial functions $P_i=x_i$ with the adjustments using the Brouwer fixed point (or simply a center of mass argument) and Borsuk–Ulam theorems already effected so that

$$(6.3)\qquad \lambda_{i+1} \le \lambda_1 + \frac{\int_\Omega u_1^2\,dx}{\int_\Omega x_i^2u_1^2\,dx} \quad \text{for } i=1,2,\dots,n$$

(this is just (5.15) with the order of the x_i's reversed) and hence, by summing,

$$(6.4)\qquad \sum_{i=1}^n \lambda_{i+1} \le n\lambda_1 + \sum_{i=1}^n \frac{\int_\Omega u_1^2\,dx}{\int_\Omega x_i^2u_1^2\,dx}.$$

One then uses integration by parts

$$-2\int_\Omega x_i u_1 u_{1x_i}dx = -\int_\Omega x_i(u_1^2)_{x_i}dx = \int_\Omega u_1^2 dx \tag{6.5}$$

and the Cauchy–Schwarz inequality

$$\left(\int_\Omega u_1^2 dx\right)^2 = 4\left(\int_\Omega x_i u_1 u_{1x_i}dx\right)^2 \le 4\left(\int_\Omega x_i^2 u_1^2 dx\right)\left(\int_\Omega u_{1x_i}^2 dx\right) \tag{6.6}$$

to arrive at

$$\frac{\int_\Omega u_1^2 dx}{\int_\Omega x_i^2 u_1^2 dx} \le \frac{4\int_\Omega u_{1x_i}^2 dx}{\int_\Omega u_1^2 dx}. \tag{6.7}$$

Substituting this into (6.4) yields

$$\lambda_2 + \cdots + \lambda_{n+1} \le n\lambda_1 + 4\frac{\int_\Omega |\nabla u_1|^2 dx}{\int_\Omega u_1^2 dx} = (n+4)\lambda_1, \tag{6.8}$$

and (6.2) is proved. This is essentially the route used by Payne, Pólya, and Weinberger in the two-dimensional case except that rather than obtaining the individual inequalities (6.3) they used the extended Rayleigh–Ritz principle to bound the sum $\lambda_2 + \lambda_3$ by the trace of a 2×2 matrix with matrix elements $(\varphi_i, -\Delta\varphi_j)$, where the φ_i's are chosen to be orthonormal and also orthogonal to u_1. Of course, their φ_i's are nothing but $x_i u_1$, and the required orthogonality conditions are obtained via translation and rotation of the coordinate axes as in the argument above.

Similarly, the bound

$$(\lambda_2 + \lambda_3)/\lambda_1 \le 5 + \lambda_1/\lambda_2 \tag{6.9}$$

of Brands [12] extends easily to n dimensions as

$$(\lambda_2 + \lambda_3 + \cdots + \lambda_{n+1})/\lambda_1 \le n + 3 + \lambda_1/\lambda_2. \tag{6.10}$$

To obtain this one follows Brands or PPW (as given above) to (6.4), but then one estimates the sum on the right-hand side via

$$\sum_{i=1}^{n} \frac{\int_\Omega u_1^2 dx}{\int_\Omega x_i^2 u_1^2 dx} \le \frac{(\alpha+1)^2}{2\alpha - 1}\lambda_1 A(\alpha), \tag{6.11}$$

where

$$A(\alpha) = \frac{\left(\int_\Omega u_1^{2\alpha} dx\right)\left(\int_\Omega u_1^2 dx\right)}{\left(\int_\Omega u_1^{\alpha+1} dx\right)^2} \tag{6.12}$$

with α a parameter larger than $\frac{1}{2}$ (setting $\alpha = 1$ recovers the argument of PPW). Brands then finds, with $\nu \equiv \lambda_2/\lambda_1$, that

$$A(\alpha) \le \frac{(2\alpha - 1)(\nu - 1)}{(2\alpha - 1)\nu - \alpha^2} \tag{6.13}$$

if $1 < \alpha < \nu + \sqrt{\nu^2 - \nu}$. Therefore we have

$$(\lambda_2 + \lambda_3 + \cdots + \lambda_{n+1})/\lambda_1 \le n + \frac{(\alpha+1)^2(\nu-1)}{(2\alpha-1)\nu - \alpha^2}, \tag{6.14}$$

and since the minimum of the right-hand side occurs at $\alpha = 2\nu/(\nu+1)$ (which is between 1 and $\nu + \sqrt{\nu^2 - \nu}$) (6.10) follows. It is clear from their remarks [23, p. 538] that Hile and Protter knew of the generalization (6.10) of Brands' bound (6.9).

We turn now to deriving the consequences of the Brands bound (6.10) for λ_3/λ_1 and $(\lambda_2 + \lambda_3)/\lambda_1$. We obtain the following theorem.

THEOREM 6.1. *The eigenvalue ratios λ_3/λ_1 and $(\lambda_2+\lambda_3)/\lambda_1$ for the n-dimensional Dirichlet Laplacian on a bounded domain $\Omega \subset \mathbb{R}^n$ $(n \geq 2)$ obey*

$$\lambda_3/\lambda_1 \leq \frac{2}{n} + \frac{(n+2)\left(n^2+n+2+\sqrt{n^4+6n^3+13n^2-4n+4}\right)}{2n(n^2+2n-2)} \tag{6.15}$$

and, for $n \geq 3$,

$$(\lambda_2+\lambda_3)/\lambda_1 \leq 1 + \frac{3+\sqrt{n^2+10n+9}}{n}. \tag{6.16}$$

Proof. Brands' bound (6.10) yields

$$(\lambda_2+(n-1)\lambda_3)/\lambda_1 \leq n+3+\lambda_1/\lambda_2, \tag{6.17}$$

which, with $x \equiv \lambda_2/\lambda_1$ and $y \equiv \lambda_3/\lambda_1$, becomes

$$y \leq (n+3+1/x-x)/(n-1). \tag{6.18}$$

We shall use this bound in conjunction with the bound $\lambda_3 - \lambda_2 \leq 2(\lambda_1+\lambda_2)/n$ of Payne, Pólya, and Weinberger [35], which, in our new notation, reads

$$y \leq \frac{2}{n} + \left(1+\frac{2}{n}\right)x. \tag{6.19}$$

(While this bound has been improved upon by Hile and Protter [23] their bound would lead to much additional complication; (6.19) is more than adequate for our purposes here.) We shall examine what the region in the xy-plane satisfying (6.18) and (6.19), along with the two trivial inequalities $x > 1$ and $y \geq x$, says about λ_3/λ_1 and $(\lambda_2+\lambda_3)/\lambda_1$. First we observe that since the right-hand side of (6.18) is decreasing in x while that of (6.19) is increasing, the maximum value of y for this region is the y corresponding to the intersection of the right-hand sides. The point (x_0, y_0) of intersection has

$$x_0 = \frac{n^2+n+2+\sqrt{n^4+6n^3+13n^2-4n+4}}{2(n^2+2n-2)}, \tag{6.20}$$

and so

$$\frac{\lambda_3}{\lambda_1} = y \leq y_0 = \frac{2}{n} + \frac{n+2}{n}x_0 \tag{6.21}$$

for all points in the region, i.e., (6.15) holds. Note that the $n=2$ case of this gives the bound $\lambda_3/\lambda_1 \leq (7+2\sqrt{7})/3 \approx 4.097$, found originally by Brands.

For (6.16) we begin by observing that the largest value of x in the region occurs at the intersection of (6.18) with $y = x$. Solving for x yields the value

$$x_1 = \left(n+3+\sqrt{n^2+10n+9}\right)/2n, \tag{6.22}$$

so that we obtain the bound

$$\frac{\lambda_2}{\lambda_1} \leq \frac{n+3+\sqrt{n^2+10n+9}}{2n}. \tag{6.23}$$

This bound was derived earlier from Brands' work by Hile and Protter [23, p. 538]. Next observe from (6.18) that

$$x + y \leq [n + 3 + 1/x + (n - 2)x]/(n - 1) \tag{6.24}$$

and that to find an upper bound to $x + y$ on the region of interest we only need to find the maximum of the right-hand side for $x_0 \leq x \leq x_1$. This latter fact is obvious from the fact that, aside from intersections with the boundary determined by (6.18), none of the other boundary curves $x = 1, y = x$, or $y = x + 2(1 + x)/n$ can contain a point in the region with maximal value of $x + y$. Since the right-hand side of (6.24) has a minimum at $x = 1/\sqrt{n-2}$ (for $n \geq 3$; for $n = 2$ it is decreasing for all $x > 0$) and this number is less than or equal to 1 for $n \geq 3$ it is clear that for $n \geq 3$ the right-hand side of (6.24) is increasing for all x's of interest. Thus, an upper bound for $(\lambda_2 + \lambda_3)/\lambda_1$ will be given by evaluating the right-hand side of (6.24) at x_1 and, since at the point of our region that this singles out we also have $y = x$, this leads to $(\lambda_2 + \lambda_3)/\lambda_1 = x + y \leq 2x_1 = 1 + (3 + \sqrt{n^2 + 10n + 9})/n$, which is (6.16). □

Remarks. (1) If $n = 2$, then the completion of the argument for $x+y$ would yield $(\lambda_2 + \lambda_3)/\lambda_1 = x + y \leq x_0 + y_0 = 3 + \sqrt{7} \approx 5.646$, the result found earlier by Brands.

(2) Our result (6.16) shows that the result (6.23) of Hile and Protter found by generalizing Brands' argument to n dimensions is actually true in the stronger form that twice the bound on λ_2/λ_1 is in fact a bound on $(\lambda_2 + \lambda_3)/\lambda_1$, except perhaps in two dimensions. This same duplication effect and more also holds for the PPW upper bound of $1 + 4/n$ even beginning at two dimensions, i.e., $(\lambda_2 + \cdots + \lambda_{k+1})/\lambda_1 \leq k(1 + 4/n)$ for $n \geq 2$ and $1 \leq k \leq n$. In two dimensions Brands' argument yields only $(\lambda_2 + \lambda_3)/\lambda_1 \leq 5.646$, whereas for λ_2/λ_1 he obtains $\lambda_2/\lambda_1 \leq (5 + \sqrt{33})/4 \approx 2.686$, and $2(2.686) \approx 5.372 < 5.646$.

(3) Arguments similar to those in our proof above can be made by replacing the bound (6.17) with the bound

$$(\lambda_2 + (n - 1)\lambda_3)/\lambda_1 \leq n + 4, \tag{6.25}$$

which derives from the PPW bound (6.2). The analogous, but slightly weaker, results are then

$$\lambda_3/\lambda_1 \leq \frac{2}{n} + \frac{(n+2)(n^2 + 2n + 2)}{n(n^2 + 2n - 2)} \quad \text{for } n \geq 2 \tag{6.26}$$

and

$$(\lambda_2 + \lambda_3)/\lambda_1 \leq 2(1 + 4/n) \quad \text{for } n \geq 2. \tag{6.27}$$

(The second of these, of course, follows directly from the PPW bound (6.2) and even generalizes as in the previous remark.)

(4) The bound (6.16) could even be improved further by including the inequality $\lambda_2/\lambda_1 = x \leq \left(j_{n/2,1}/j_{n/2-1,1}\right)^2 \equiv K_n$ (from our proof of the Payne–Pólya–Weinberger conjecture [5], [6]) as a boundary curve for our region. This gives

$$(\lambda_2 + \lambda_3)/\lambda_1 = x + y \leq [n + 3 + K_n^{-1} + (n - 2)K_n]/(n - 1) \tag{6.28}$$

for $n \geq 3$ in place of (6.16) (with (6.25) as in Remark (3) we would obtain only $[n+4+(n-2)K_n]/(n-1)$ as the upper bound). Similarly, by taking into account the inequality

$$y \leq 1 + \frac{(n-1)(x-1)}{C_n(x-1)-1} \quad \text{for } x > 1 + C_n^{-1}, \tag{6.29}$$

where $C_n \equiv [2j_{n/2-1,1}^2 + n(n-4)]/6$, which derives from (5.13) above, and finding the point of intersection of its right-hand side with (6.17) or (6.25) and with the boundary curve $y = x$ or $x = K_n$, we could get further improvements, though at the expense of more complication (see Theorems 6.2 and 6.3 below).

We now turn to comparison of the large n asymptotics of our various upper bounds for λ_3/λ_1 and $(\lambda_2+\lambda_3)/\lambda_1$. We begin with λ_3/λ_1. One finds that the right-hand side of (6.15) behaves like

$$1 + \frac{4}{n} + \frac{4}{n^2} - \frac{4}{n^3} + O(n^{-4}), \tag{6.30}$$

while that of (6.26) behaves like

$$1 + \frac{4}{n} + \frac{4}{n^2} + O(n^{-4}). \tag{6.31}$$

These bear comparison to the large n asymptotics of the bounds on λ_2/λ_1 found here and in the previous section. In order of accuracy we have the bounds $1+4/n$ (Payne, Pólya, and Weinberger/Thompson), $1 + 4/n - 4/n^2 + 20/n^3 + O(1/n^4)$ (Brands as generalized by Hile and Protter (6.23)), $1 + 4/n - c/n^{5/3} + 16/n^2 + O(n^{-7/3})$ (Chiti, as evaluated by us (5.3)), and $1 + 4/n - c/n^{5/3} + 12/n^2 + O(n^{-7/3})$ (Ashbaugh and Benguria (1.1)). Here $c = 2^{10/3}|a_1|/3 \approx 7.85555$, where a_1 = first negative zero of the Airy function $Ai(x)$ as discussed in §5 (for even more details, see our discussion following (6.53) below).

At this point we could obtain, via combination of our separate results for λ_3/λ_1 and λ_2/λ_1, a bound for $(\lambda_2+\lambda_3)/\lambda_1$ having asymptotic form $2 + 8/n - c/n^{5/3} + 16/n^2 + O(n^{-7/3}) = 2(1 + 4/n - c/2n^{5/3} + 8/n^2) + O(n^{-7/3})$. Using only bounds deriving from Brands' work we would get only $2 + 8/n + 16/n^3 + O(1/n^4) = 2(1 + 4/n + 8/n^3) + O(1/n^4)$. However, we can get better bounds out of our work above since it dealt directly with the combination $(\lambda_2+\lambda_3)/\lambda_1$. In particular, (6.16) behaves like $2 + 8/n - 8/n^2 + 40/n^3 + O(1/n^4) = 2(1 + 4/n - 4/n^2 + 20/n^3) + O(1/n^4)$, which gives a better result than combining the separate Brands-type bounds for λ_2/λ_1 and λ_3/λ_1, though it still falls short of the result which made use of our bound $\lambda_2/\lambda_1 \leq (j_{n/2,1}/j_{n/2-1,1})^2$, at least asymptotically. To beat the bound that goes as $2+8/n-c/n^{5/3}+16/n^2$ we need to investigate the asymptotics of the bound in (6.28). This gives the somewhat better asymptotic form

$$2 + 8/n - c/n^{5/3} + 8/n^2 + O(n^{-7/3}). \tag{6.32}$$

While this is still not the last word on the large n asymptotics of bounds for $(\lambda_2 + \lambda_3)/\lambda_1$, we defer further discussion until after our next two theorems. These theorems take into account the additional bound (6.29). We give these theorems in a form that would be of use for any dimension n and also in their asymptotic form for large n.

THEOREM 6.2. *With hypotheses as given previously, the eigenvalue ratio* $(\lambda_2 + \lambda_3)/\lambda_1$ *obeys the bound*

(6.33)

$$(\lambda_2+\lambda_3)/\lambda_1 \le \max\{[n+4+(n-2)x_1']/(n-1),\ 1+K_n+(n-1)(K_n-1)/[C_n(K_n-1)-1]\},$$

where $C_n = [2j_{n/2-1,1}^2 + n(n-4)]/6$, $K_n = (j_{n/2,1}/j_{n/2-1,1})^2$, *and*

$$x_1' = 1 + \{4C_n - n^2 + 2n - [(4C_n - n^2 + 2n)^2 - 16C_n]^{1/2}\}/2C_n. \tag{6.34}$$

Moreover, for sufficiently large n the second expression on the right-hand side of (6.33) *prevails, and hence*

$$(\lambda_2 + \lambda_3)/\lambda_1 \le 2 + 8/n - 2c/n^{5/3} + 28/n^2 + O(n^{-7/3}) \tag{6.35}$$

for large n.

Note. If it occurs that $x_1' > K_n$, then (6.33) can be replaced by $(\lambda_2 + \lambda_3)/\lambda_1 \le [n + 4 + (n-2)K_n]/(n-1)$. However, we do not think this possibility ever actually occurs, so we have chosen not to clutter the theorem with additional cases. The theorem is, of course, correct without this extra case.

Proof. The approach is as in the proof of Theorem 6.1. We consider the region determined by the inequalities $x > 1$, $y \ge x$, (6.19), (6.25), (6.29), and $x \le K_n$ and show that the maximum of $x + y$ must occur at one of two vertices of this region. These two vertices are (1) the vertex where the boundary curves determined by (6.25) and (6.29) meet and (2) the vertex where the boundary curves determined by (6.29) and $x \le K_n$ meet. To see this one observes that the curves

$$y = 1 + \frac{(n-1)(x-1)}{C_n(x-1)-1} \quad \text{for } x > 1 + C_n^{-1} \tag{6.36}$$

and

$$y = \frac{n+4-x}{n-1} \tag{6.37}$$

have points of intersection for

$$x_\pm = \{6C_n - n^2 + 2n \pm [(4C_n - n^2 + 2n)^2 - 16C_n]^{1/2}\}/2C_n.$$

To see that these roots are real one uses a recent result of Lorch [26] which says that Chiti's bound for λ_2/λ_1, $1 + n/C_n$, is less than the PPW/Thompson bound, $1 + 4/n$, for all n (i.e., $C_n > n^2/4$ for all $n = 1, 2, 3, \ldots$; cf. (5.5) and (5.6)). From this it follows that (6.37) intersects $y = x$ farther out along $y = x$ than does (6.36), and then the asymptotics of the two curves tell us that they must cross at the values $x_\pm$ as determined above. Hence the curve determined by (6.36) lies above that determined by (6.37) for $1 + C_n^{-1} < x < x_-$ and for $x > x_+$ while the opposite is true on the intermediate interval $x_- < x < x_+$. To continue it is enough to observe as before that the maximum value of $x + y$ attained on the allowed region must occur either along (6.36) or (6.37). Since (6.33) yields the PPW bound of 6 for $n = 2$ we can assume henceforth that $n > 2$. From (6.36) and (6.37) we define the related functions

$$F(x) = x + \frac{n+4-x}{n-1} = \frac{n+4+(n-2)x}{n-1} \tag{6.38}$$

and

$$G(x) = 1 + x + \frac{(n-1)(x-1)}{C_n(x-1)-1}. \tag{6.39}$$

As in our previous proof we note that $F(x)$ is an increasing function. Hence the maximum of $x+y$ on the interval $[1, x_-]$ must be $F(x_-)$ which is the first expression found in our bound (6.33) (even if $x_- > K_n$ this expression will serve as an upper bound).

For the other part of the bound we concentrate on $G(x)$ on the interval $[x_-, K_n]$ (where we admit for the time being the possibility that this set vanishes if $x_- > K_n$ in which case the second expression on the right in (6.33) can be dropped). We compute easily that $G'(x) = 0$ at

$$x_2 = 1 + C_n^{-1}(1 + \sqrt{n-1}) \tag{6.40}$$

and that this gives a relative minimum with G decreasing on $(1 + C_n^{-1}, x_2]$ and increasing on $[x_2, \infty)$. Thus, no matter where x_2 is located with respect to K_n the maximum value of $G(x)$ (hence $x+y$) on the interval $[x_-, K_n]$ is given by $\max\{G(x_-), G(K_n)\}$, and this is the conclusion of the theorem (note that $F(x_1') = G(x_1')$ since $x_1' = x_-$ gives a point of intersection of the graphs of F and G).

The proof thus far is a little unsatisfying in that we haven't pinned down where x_-, x_+, x_2, and K_n lie with respect to each other and hence the thought may linger that our bound could be improved if we were to investigate the locations of these points in greater detail. We now do this in the large n regime, where we show that $x_2 < x_- < K_n < x_+$, and hence that, at least for large n, the bound $G(K_n)$ is the best to be had within the context of our argument. (That $K_n < x_+$ for all n already follows from the result of Lorch used above.)

Asymptotic formulas for x_2, x_-, K_n, and x_+ are as follows:

$$x_2 = 1 + \frac{4}{n^{3/2}} + \frac{4}{n^2} + O(n^{-5/2}), \tag{6.41}$$

$$x_- = 1 + \frac{16}{cn^{4/3}} + \frac{128}{c^2 n^{5/3}} + O(n^{-2}), \tag{6.42}$$

$$K_n = 1 + \frac{4}{n} - \frac{c}{n^{5/3}} + \frac{12}{n^2} + O(n^{-7/3}), \tag{6.43}$$

and

$$x_+ = 1 + \frac{c}{n^{2/3}} - \frac{8}{n} + O(n^{-4/3}). \tag{6.44}$$

These show that the relevant bound for $(\lambda_2 + \lambda_3)/\lambda_1$ in the large n regime is $G(K_n)$ and asymptotically we have

$$G(K_n) = 2 + \frac{8}{n} - \frac{2c}{n^{5/3}} + \frac{28}{n^2} + O(n^{-7/3}),$$

which proves the last part of our theorem. □

Remark. One could add to the statement of the theorem the fact that if $x_2 \leq x_1'(= x_-)$ then one has immediately the bound $(\lambda_2 + \lambda_3)/\lambda_1 \leq G(K_n)$.

A slight, additional improvement to Theorem 6.2 can be obtained by using the bound (6.18) instead of (6.25) to define the allowed region. When solving for the intersection points of $G(x)$ and $\tilde{F}(x) \equiv [n+3+x^{-1}+(n-2)x]/(n-1)$ one encounters a cubic of which the second and third roots play the roles of x_- and x_+, respectively (the first root is less than $1+C_n^{-1}$ and, therefore, is of no relevance here). Specifically, one encounters the cubic

$$C_n(x-1)^3 - (2C_n - n^2 + 2n)(x-1)^2 - (4C_n - n^2 + 2n - 3)(x-1) + 4 = 0, \tag{6.45}$$

and since $C_n = n^2/4 + cn^{4/3}/16 - n + O(n^{2/3})$ the roots r_1, r_2, and r_3 $(r_1 \leq r_2 \leq r_3)$ behave asymptotically for large n as

$$r_1 \sim -1, \quad r_2 \sim 1 + 16/cn^{4/3}, \quad r_3 \sim 1 + c/2n^{2/3}.$$

That these three roots are real for all $n = 1, 2, 3, \ldots$ follows from a second result of Lorch [26] which says that Chiti's bound for λ_2/λ_1 is less than the Brands/Hile–Protter bound as given in (6.23) above, i.e., that $1 + n/C_n < (n + 3 + \sqrt{n^2 + 10n + 9})/2n$ for $n = 1, 2, 3, \ldots$. Just as Lorch's first result could be stated as the lower bound (5.6) for the Bessel function zero $j_{p,1}$, we can state this result as

$$j_{p,1}^2 > (p+1)\left[3\sqrt{4p^2 + 28p + 33} - 2p + 5\right]/4.$$

Again, this is valid for all $p > -1$ though we only use it for $p = n/2 - 1$ with $n = 2, 3, 4, \ldots$. Conclusions analogous to Theorem 6.2 can now be drawn by following the same approach (again, asymptotically it is $G(K_n)$ that matters, not $G(r_2)$ $(= \tilde{F}(r_2))$, and therefore the asymptotic result is that of (6.35) in Theorem 6.2 and will not be repeated). This proves the following.

THEOREM 6.3. *With hypotheses and notation as before, for $n \geq 3$ the eigenvalue ratio $(\lambda_2 + \lambda_3)/\lambda_1$ obeys the bound*

$$(\lambda_2 + \lambda_3)/\lambda_1 \leq \max\{[n + 3 + r_2^{-1} + (n-2)r_2]/(n-1),\ G(K_n)\}, \tag{6.46}$$

where r_2 is the intermediate root of the cubic (6.45).

Remark. For $n = 2$ our approach above yields again the result of Brands $(\lambda_2 + \lambda_3)/\lambda_1 \leq 3 + \sqrt{7}$. This is because in that one case $\tilde{F}(x)$ is decreasing across the entire interval of interest instead of increasing as happens for all other cases. (Note that when $n = 2$, $F(x)$ is constantly 6. This explains why the bound in Theorem 6.2 worked even for $n = 2$.)

The asymptotic expansion for our bound (see (6.35)) begins like that of $2K_n$, the conjectured best possible bound. This can be taken as additional evidence in support of this conjecture. Moreover, if one replaces the bound $G(K_n)$ by $G(1+n/C_n)$ (which is certainly permissible for large n since G will then be increasing over the interval of interest and $K_n < 1 + n/C_n$ is true for all n; $1 + n/C_n$ here is the constant occurring in Chiti's bound for λ_2/λ_1), then one gets $(\lambda_2 + \lambda_3)/\lambda_1 \leq 2(1 + n/C_n)$. That is, Chiti's bound has the duplication property for n sufficiently large that $G(1 + n/C_n)$ is greater than or equal to either of $G(x_-)$ or $G(r_2)$. The asymptotic expansion of $2(1 + n/C_n)$ differs from (6.35) only in terms $O(n^{-2})$ and beyond; in particular (see

(5.4)), the term $28/n^2$ in (6.35) gets replaced by $32/n^2$, while the conjectured best bound has $24/n^2$ (see (5.1)).

If one applies the arguments we use in this section to the conjectured bound $(\lambda_2 + \cdots + \lambda_{n+1})/\lambda_1 \le nK_n$, then $(\lambda_2 + \lambda_3)/\lambda_1 \le 2K_n$ of course follows as well as

$$\begin{aligned} \lambda_3/\lambda_1 \le & \frac{2}{n} + \frac{(n+2)[n^2K_n - 2(n-1)]}{n(n^2+2n-2)} \\ = & 1 + \frac{4}{n} - \frac{c}{n^{5/3}} + \frac{16}{n^2} + \frac{k_1}{n^{7/3}} - \frac{17c}{3n^{8/3}} + \frac{k_2}{n^3} + O(n^{-10/3}). \end{aligned} \tag{6.47}$$

Here the constant c has already been defined, while k_1 and k_2 are positive constants which will be explained more fully in connection with (6.52) and (6.53) below. A comparison with (6.30) (or (6.31)) shows that the bound (6.15) that we have proved here is only slightly worse than this for large n (cf. also (6.55) below).

There is one more improvement that we can make to some of the results in this section. That is to incorporate the bound $\lambda_3/\lambda_2 < K_n$ (i.e., $y < K_n x$) which was recently proved by us [7] (see also [8]) into our constraints on the allowed region. It turns out that this bound effectively supersedes the bound (6.19) of Payne, Pólya, and Weinberger at least in the matter of bounding λ_3/λ_1 for large dimension n (and probably for all $n = 2, 3, 4, \dots$). We begin by examining the intersection point of the lines $y = K_n x$ and $y = 2/n + (1 + 2/n)x$. This point is

$$\overline{x}_0 = \frac{2}{nK_n - (n+2)} = 1 + \frac{c}{2n^{2/3}} - \frac{6}{n} + O(n^{-4/3}). \tag{6.48}$$

At least asymptotically, this point is larger than

$$x_0 = \{n^2 + n + 2 + [n^4 + 6n^3 + 13n^2 - 4n + 4]^{1/2}\}/2(n^2 + 2n - 2),$$

corresponding to the point of intersection of $y = 2/n + (1 + 2/n)x$ and $y = (n + 3 + x^{-1} - x)/(n-1)$, which behaves as $1 + 4/n^2 - 12/n^3 + O(n^{-4})$. Since at $x = 1$ $K_n x < 2/n + (1+2/n)x$, this means that $K_n x$ is the controlling bound in the region of interest for $\max \lambda_3/\lambda_1$. Thus we go back and find the points of intersection of $y = K_n x$ with $y = (n + 3 + x^{-1} - x)/(n-1)$ and $y = (n + 4 - x)/(n-1)$, and proceeding in the now familiar way we arrive at the following theorem.

THEOREM 6.4. *With hypotheses and notation as above the ratio* λ_3/λ_1 *obeys*

$$\lambda_3/\lambda_1 < K_n \frac{n + 3 + [(n+3)^2 + 4(n-1)K_n + 4]^{1/2}}{2(n-1)K_n + 2}. \tag{6.49}$$

This is a better bound than (6.15) *if and only if* $x_0 < \overline{x}_0$, *where* $\overline{x}_0 = 2/[nK_n - (n+2)]$ *and* x_0 *is given by* (6.20). *In particular,* (6.49) *is better than* (6.15) *for sufficiently large* n.

Proof. The relevant point of intersection of $y = K_n x$ with $y = (n + 3 + x^{-1} - x)/(n-1)$ is easily found to be

$$x_0' = \frac{n + 3 + [(n+3)^2 + 4(n-1)K_n + 4]^{1/2}}{2(n-1)K_n + 2} \tag{6.50}$$

and that with $y = (n + 4 - x)/(n-1)$ is

$$x_0'' = \frac{n+4}{(n-1)K_n + 1}. \tag{6.51}$$

These have the asymptotic expansions

$$x_0' = 1 + \frac{c}{n^{5/3}} - \frac{8}{n^2} - \frac{k_1}{n^{7/3}} - \frac{c}{3n^{8/3}} + \frac{52 - k_2}{n^3} + O(n^{-10/3}) \tag{6.52}$$

and

$$x_0'' = 1 + \frac{c}{n^{5/3}} - \frac{8}{n^2} - \frac{k_1}{n^{7/3}} + \frac{2c}{3n^{8/3}} + \frac{44 - k_2}{n^3} + O(n^{-10/3}), \tag{6.53}$$

where c is a positive constant as defined above and k_1 and k_2 are higher coefficients in the asymptotic expansion of K_n; specifically, we have $K_n = 1 + 4/n - c/n^{5/3} + 12/n^2 + k_1/n^{7/3} - 17c/3n^{8/3} + k_2/n^3 + O(n^{-10/3})$ with k_1 and k_2 being positive constants with values $k_1 \approx 9.256$, $k_2 \approx 25.31$, which are determined from the coefficients in the asymptotic expansion of the Bessel function zero $j_{p,1}$ for large p as given by Abramowitz and Stegun [1, p. 371, eq. 9.5.14]. (These coefficients can all be related to the first negative zero $a_1 \approx -2.33810741$ of the Airy function $Ai(x)$; in particular, $c = -2^{10/3}a_1/3 \approx 7.85555$, $k_1 = 2^{14/3}a_1^2/15$, and $k_2 = 16(16a_1^3 + 1035)/525$.) Equations (6.52) and (6.53) show already that asymptotically $x_0 < x_0' < x_0'' < \bar{x}_0$. (It is also clear that for sufficiently large n, x_0'' is less than both x_- and r_2 so the bound $y < K_n x$ will not give improvements to the bounds for $(\lambda_2 + \lambda_3)/\lambda_1$ found previously, at least for large n.) By the same arguments used previously one now has $\lambda_3/\lambda_1 < K_n x_0'$ and $K_n x_0''$, with $K_n x_0'$ being a slightly better bound (for all n; the only advantage that x_0'' has to recommend it over x_0' is its relative simplicity). □

Remarks. (1) The condition for when (6.49) is better than (6.15) can be given in other equivalent terms. We have $x_0 < \bar{x}_0 \Leftrightarrow x_0' < \bar{x}_0 \Leftrightarrow x_0 < x_0'$ (so that either $x_0 < x_0' < \bar{x}_0$, $\bar{x}_0 < x_0' < x_0$, or $x_0 = x_0' = \bar{x}_0$).

(2) The quantity $K_n x_0'$ (the bound for λ_3/λ_1 in (6.49)) has asymptotic behavior

$$1 + \frac{4}{n} + \frac{4}{n^2} - \frac{2c}{n^{8/3}} + \frac{20}{n^3} + O(n^{-10/3}), \tag{6.54}$$

which is slightly better than the asymptotic form (6.30) found for our best previous bound.

(3) If one uses $y < K_n x$ to improve the conjectured result (6.47) for the bound on λ_3/λ_1 that would follow from the conjecture $(\lambda_2 + \cdots + \lambda_{n+1})/\lambda_1 \le nK_n$, one obtains

(6.55)

$$\lambda_3/\lambda_1 \le \frac{nK_n^2}{(n-1)K_n + 1} = 1 + \frac{4}{n} - \frac{c}{n^{5/3}} + \frac{16}{n^2} + \frac{k_1}{n^{7/3}} - \frac{20c}{3n^{8/3}} + \frac{k_2 + 12}{n^3} + O(n^{-10/3}).$$

As can be seen, this would be marginally better than (6.47). Equations (6.47) and (6.55) give us something to shoot for in trying to improve upon (6.49) and its asymptotic form (6.54), though it is by no means clear that even (6.55) could not be improved upon.

It might also be noted that our best bounds on $(\lambda_2 + \lambda_3)/\lambda_1$ for large n, at least, depended on the inequality (5.13) through $y \le 1 + (n-1)(x-1)/[C_n(x-1) - 1]$. One can consider the consequences if (5.13) could be replaced by the stronger inequality (5.16) which is itself weaker than (5.19) (see our discussion in §5). This would amount to replacing the constant C_n by $n/(K_n - 1)$ in the operative inequalities and would imply that $(\lambda_2 + \lambda_3)/\lambda_1 \le 2K_n$ for all sufficiently large n. Thus the relatively weak but optimal conjectured inequality (5.16) would imply the best possible bound on $(\lambda_2 + \lambda_3)/\lambda_1$ for large dimensions (perhaps this approach could be made to yield $(\lambda_2 + \lambda_3)/\lambda_1 \leq 2K_n$ for all dimensions higher than three; it cannot work in two dimensions because $5 + 1/x$ is decreasing and also x_2, the point where $G(x)$ reaches its minimum, is larger than K_2).

We conclude by briefly summarizing what we have accomplished in this section and by applying our work to obtain numerical bounds in the three- and four-dimensional cases. We also make some final observations concerning the two-dimensional case. For $n \geq 2$ the boundary curves that matter for λ_3/λ_1 are $y = [n + 3 + x^{-1} - x]/(n - 1)$ and whichever of $y = K_n x$ or $y = 2/n + (1 + 2/n)x$ gives a lower intersection point with the first. This is because the first function is decreasing while the latter two are increasing. Theorems 6.1 and 6.4 cover the two possibilities. For $(\lambda_2 + \lambda_3)/\lambda_1$ the boundary curves that matter for $n \geq 3$ are $x+y = [n+3+x^{-1}+(n-2)x]/(n-1)$ again, $x+y = 1+x+(n-1)(x-1)/[C_n(x-1)-1]$, and $x = K_n$, and hence the bounds that we obtain are the values of $x+y$ corresponding to certain points of intersection of these curves (at most two of these points of intersection ever matter at the same time). The relevant facts are now that the right-hand side of the first equation is an increasing function of x (for $n \geq 3$ and $x \geq 1$) while that of the second is convex. Thus if the second equation matters its endpoint values over the interval on which it matters determine an upper bound for $(\lambda_2 + \lambda_3)/\lambda_1$ and if it does not, then $[n+3+K_n^{-1}+(n-2)K_n]/(n-1)$ is an upper bound. This information is included in (6.28) and Theorem 6.3. In all cases for both λ_3/λ_1 and $(\lambda_2 + \lambda_3)/\lambda_1$, except that of $(\lambda_2 + \lambda_3)/\lambda_1$ when $n = 2$, the boundary curve $y = [n + 4 - x]/(n - 1)$ can replace $y = [n+3+x^{-1} - x]/(n - 1)$ without changing this qualitative picture; it will give slightly worse results but the expressions giving them may be somewhat simpler. This explains the origins of our results concerning $(\lambda_2 + \lambda_3)/\lambda_1$ in Theorems 6.1 (which used only $y = x$ and not $x = K_n$ as one of the boundary curves) and 6.2.

It may in fact be true that of all the bounds derived above for λ_3/λ_1, Theorem 6.4 is always the best for $n \geq 2$ and that for $(\lambda_2 + \lambda_3)/\lambda_1$, $G(K_n)$ is always the best bound for $n \geq 4$. Certainly for large enough n this is what happens. It only remains to determine at what value of n this occurs. It is because this matter is not easy to settle that we stated our theorems in terms of the various contingencies that we did. However, indications from low order cases (see below for the cases $n = 2, 3$, and 4) are that already by $n = 4$ these bounds predominate.

For $n = 3$ our work above gives the numerical results

$$\lambda_3/\lambda_1 < \frac{K_3(3 + \sqrt{2K_3 + 10})}{2K_3 + 1} \approx 2.7137 \tag{6.56}$$

and

$$(\lambda_2 + \lambda_3)/\lambda_1 \leq 4.1499. \tag{6.57}$$

We also note the bounds

$$\lambda_2/\lambda_1 \leq K_3 \approx 2.0457, \tag{6.58}$$

which is best possible, and

$$(\lambda_2 + \lambda_3 + \lambda_4)/\lambda_1 \leq 7. \tag{6.59}$$

The conjectured best possible bounds for $(\lambda_2 + \lambda_3)/\lambda_1$ and $(\lambda_2 + \lambda_3 + \lambda_4)/\lambda_1$ are $2K_3 \approx 4.0915$ and $3K_3 \approx 6.1372$, respectively. There is no reasonable guess as to what the best bound for λ_3/λ_1 might be. The bound in (6.56) derives from the point of intersection of $y = K_3x$ and $y = (6+x^{-1}-x)/2$ while that in (6.57) comes from the intersection point of $y = (6+x^{-1}-x)/2$ and $y = 1+2(x-1)/[C_3(x-1)-1]$ which has x coordinate ≈ 1.717614 (specifically, this value of x is the middle root of the cubic $C_3x^3 - (5C_3 - 3)x^2 + 3C_3x + (C_3 + 1) = 0$ with $C_3 = (2\pi^2 - 3)/6 \approx 2.789868$). To see how close $G(K_3)$ comes to taking over in (6.46) (and hence in (6.57)) we note that $G(K_3) \approx 4.1365$, which is only slightly smaller than the value 4.1499 found in (6.57). For $n = 4$, $G(K_4) \approx 3.6209$ does in fact prevail and it is likely that this trend continues so that $G(K_n)$ prevails for all $n \geq 4$. Also for $n = 4$ one finds $\lambda_2/\lambda_1 \leq K_4 \approx 1.7964$ and $\lambda_3/\lambda_1 < K_4(7 + \sqrt{12K_4 + 53})/2(3K_4 + 1) \approx 2.1979$, the latter coming from the point of intersection of $y = K_4x$ and $y = (7 + 1/x - x)/3$.

For $n = 2$ our work in this section gives only the bounds

$$\lambda_3/\lambda_1 < K_2(5 + \sqrt{4K_2 + 29})/2(K_2 + 1) \approx 4.0381$$

(better than PPW and Brands, not as good as Hile and Protter or Marcellini) and

$$(\lambda_2 + \lambda_3)/\lambda_1 \leq (5 + \sqrt{4K_2 + 29})/2 \approx 5.6287$$

(which has the same position relative to the bounds of the others). Hile and Protter [23] and Marcellini [27] both went beyond our results by working with an additional inequality which takes the place of $y \leq 1+2x$ (as used by Brands) or $y < K_2x$ (as used in our work here). For Hile and Protter it is the bound $y \leq 1 + x + \sqrt{1 - x + x^2}$ and for Marcellini it is $y \leq \{1 + 3x + 4x^2 - [(1 + 3x + 4x^2)^2 - 58x^3 - 4x^2 + 2x]^{1/2}\}/2x$. While Hile and Protter's inequality has the n-dimensional generalization

$$y \leq \left\{(n + 2)(1 + x) + [(n + 2)^2x^2 - 2(n^2 + 4n - 4)x + (n + 2)^2]^{1/2}\right\} \Big/ 2n, \tag{6.60}$$

it seems to be less effective than $y < K_nx$ for higher dimensions than two (at least this is so for $n = 3$ and $n = 4$; it is also significantly more complicated than $y < K_nx$). Similarly, while we have obtained

$$y \leq \frac{A - [A^2 - 4n(n-1)x\{(n^2 + 7n + 11)x^2 + 2x - 1\}]^{1/2}}{2n(n-1)x}, \tag{6.61}$$

where

$$A \equiv (n^2 + 4n - 4)x^2 + n(n + 1)x + n \tag{6.62}$$

as the n-dimensional generalization of Marcellini's bound, the bounds we have developed above would appear to yield better bounds for λ_3/λ_1 in four and more dimensions (in three dimensions the generalized Marcellini bound yields $\lambda_3/\lambda_1 \leq 2(3 + \sqrt{14})/5 \approx 2.6967$, a modest improvement upon (6.56)). All bounds in two dimensions for both λ_3/λ_1 and $(\lambda_2 + \lambda_3)/\lambda_1$ correspond to the vertex point of the allowed region at the left end of the interval of validity of $y \leq 5 + x^{-1} - x$ (or $y \leq 6 - x$ for the PPW bound); when $n \geq 3$ this is still the case for λ_3/λ_1 but not for $(\lambda_2 + \lambda_3)/\lambda_1$.

We summarize the best results for λ_3/λ_1 and $(\lambda_2 + \lambda_3)/\lambda_1$ found to date in Table 1. Note that for $n \geq 3$ the bounds for $(\lambda_2 + \lambda_3)/\lambda_1$ are already in very good agreement

TABLE 1

Best upper bounds to date for λ_3/λ_1 and $(\lambda_2+\lambda_3)/\lambda_1$ in dimensions 2, 3, and 4 (rounded to 4 places beyond the decimal).

n	bound for λ_3/λ_1	bound for $(\lambda_2+\lambda_3)/\lambda_1$	$(\lambda_2+\lambda_3)/\lambda_1$ for an n-ball
2	3.9170	5.5957	5.0775
3	2.6967	4.1499	4.0915
4	2.1979	3.6209	3.5928

with the conjectured optimal bound of $2K_n$ (which is the value of $(\lambda_2+\lambda_3)/\lambda_1$ for an n-ball).

An analysis similar to that for general n above could be employed to bound λ_4/λ_1 and $(\lambda_2+\lambda_3+\lambda_4)/\lambda_1$ in three and more dimensions. One would then have to study the allowed region in xyz-space for $x \equiv \lambda_2/\lambda_1$, $y \equiv \lambda_3/\lambda_1$, and $z \equiv \lambda_4/\lambda_1$, much as we have done above for λ_2/λ_1 and λ_3/λ_1. In such an analysis one would use the natural analogs of the inequalities we used above; our recently proved bound $\lambda_4/\lambda_2 < K_n$ (see [8]) would replace $\lambda_3/\lambda_2 < K_n$.

7. Improvements to the upper bound of Singer, Wong, Yau, and Yau.

In [45], Singer, Wong, Yau, and Yau considered the problem of bounding the fundamental gap $\lambda_2-\lambda_1$ for a Schrödinger operator $-\Delta+V(x)$ acting on $L^2(\Omega)$ for bounded domains $\Omega \subset \mathbb{R}^n$ with Dirichlet boundary conditions imposed on $\partial\Omega$. They were able to find natural upper and lower bounds to $\lambda_2-\lambda_1$. For their lower bound they impose convexity on both Ω and V, while their upper bound, which is just the PPW/Thompson bound $1+4/n$ generalized to Schrödinger operators and viewed as an upper bound on the difference $\lambda_2-\lambda_1$, applies under much less stringent hypotheses. In its most general form their upper bound reads

$$\lambda_2-\lambda_1 \le \frac{4}{n}(\lambda_1-m) \le \frac{4}{n}(\mu_1+M-m), \tag{7.1}$$

where m and M are lower and upper bounds for the potential V, i.e., $m \le V(x) \le M$ for all $x \in \Omega$, and μ_1 represents the first Dirichlet eigenvalue of $-\Delta$ on Ω. This bound follows immediately by observing that the original argument of Payne, Pólya, and Weinberger carries over essentially unchanged to Schrödinger operators with $V \ge 0$ on Ω. Thus the operator $-\Delta+V-m$ has eigenvalues λ_1-m and λ_2-m satisfying

$$(\lambda_2-m) \le \left(1+\frac{4}{n}\right)(\lambda_1-m)$$

so that the first part of (7.1) holds. To improve upon (7.1) it suffices to note (see Theorem 4.2 of [6]) that our upper bound for λ_2/λ_1 also carries over intact to Schrödinger operators with nonnegative potentials. We therefore obtain

$$\lambda_2-\lambda_1 \le \left[(j_{n/2,1}/j_{n/2-1,1})^2-1\right](\lambda_1-m) \le \left[(j_{n/2,1}/j_{n/2-1,1})^2-1\right](\mu_1+M-m) \tag{7.2}$$

as a better upper bound than (7.1). In two dimensions, for example, the multiplicative factor 2 from (7.1) is improved to approximately 1.539.

One can obtain more explicit upper bounds by using estimates for μ_1. One such bound is simply

$$\mu_1 \le 4j_{n/2-1,1}^2/D^2, \tag{7.3}$$

where D is the diameter of the largest ball that can be inscribed within Ω. As noted by Singer, Wong, Yau, and Yau, Cheng [15] gave the further bound

$$\mu_1 \leq 4j^2_{n/2-1,1}/D^2 < n^2\pi^2/D^2. \tag{7.4}$$

This comes from considering the first eigenvalue of the n-cube inscribed in the ball of diameter D, which has sides of length $D/\sqrt{n}$. Certainly for large n this bound loses a lot with respect to (7.3), specifically, $n^2\pi^2/D^2$ is worse by a factor of π^2 as n goes to infinity. To do somewhat better (for $n \geq 2$) one can employ the inequality

$$j^2_{p,1} < 2(p+1)(p+3) \tag{7.5}$$

(see Watson [48, p. 486, ineq. (5)]). This gives

$$\mu_1 < 2n(n+4)/D^2, \tag{7.6}$$

which is still high by a factor of 2 in the large n limit. With reference to the subject matter of this paper it is perhaps worth noting that (7.5) follows for $p = -\frac{1}{2}, 0, \frac{1}{2}, 1, \ldots$ from the fact that

$$\begin{aligned} 1 + \frac{3}{n} &= [\lambda_2/\lambda_1 \text{ for the } n\text{-cube }] < [\text{ Chiti's bound for } \lambda_2/\lambda_1] \\ &= 1 + \frac{6n}{2j^2_{n/2-1,1} + n(n-4)} \end{aligned} \tag{7.7}$$

(see (5.3) in §5 above).

Beyond (7.5), there is the bound of Chambers [13]

$$j^2_{p,1} < (p+1)[p+3+2\sqrt{p+2}] \tag{7.8}$$

which is both better than (7.5) for all $p > -1$ and asymptotically correct to leading order as p goes to infinity. This yields

$$\mu_1 < n\left(n+4+2\sqrt{2n+4}\right)/D^2 \tag{7.9}$$

as an improvement to the last part of (7.4) for all $n \geq 2$.

8. Concluding remarks. As mentioned in our previous paper, our results and techniques can probably be used to obtain improved results for related problems from differential geometry and possibly for the other problems considered in the original paper of Payne, Pólya, and Weinberger.

Mention might also be made of the fact that our upper bound for λ_2/λ_1 can be used to provide lower bounds on λ_1 provided λ_2 or a lower bound for λ_2 is known. This observation is a key ingredient in finding lower bounds to λ_1 for bent tubes [10].

Improvements in the constant in Marcellini's bound [27] for λ_3/λ_1 in two dimensions,

$$\lambda_3/\lambda_1 \leq 3.917, \tag{8.1}$$

or in his inequality $(\lambda_2+\lambda_3)/\lambda_1 \leq 5.596$ might well follow from some of our techniques, our extension (5.7) of Chiti's bound, and other known results. We intend to make a more extensive investigation of bounds on λ_3/λ_1 and $(\lambda_2+\lambda_3)/\lambda_1$ in two dimensions and report on them elsewhere.

Finally, we mention that all our bounds can be extended to operators (Dirichlet Laplacians, Schrödinger operators, or general elliptic operators as discussed in §4 of [6]) on unbounded domains $\Omega \subset \mathbb{R}^n$ so long as the operator has a form core of compactly supported functions, for in that case we can approximate all eigenvalues (defined via the Min-Max principle) in terms of appropriate Rayleigh quotients of compactly supported functions and then our known bounds for bounded domains Ω apply (see our discussion on p. 412 of [4] for more details on this). For example, our results apply to the n-dimensional isotropic harmonic oscillator, $H = -\Delta + |x|^2$ in $\mathbb{R}^n$, where one can also compute explicitly that $\lambda_2/\lambda_1 = 1 + 2/n$.

Recently we have successfully applied many of the ideas used in this paper and its predecessors to the Neumann eigenvalues of $-\Delta$ (we admit $|\Omega|$, the volume of Ω, into our formulas, however). These results will be presented in [9].

Appendix A: Some integrals of Bessel functions. We present here the antidifferentiation formula necessary for the explicit evaluation of the integral occurring in (5.2) and certain closely related antidifferentiation formulas.

We have, where $Z_p(x)$ represents any solution to Bessel's equation of order p,

$$\int xZ_p(x)^2dx = \frac{1}{2}(x^2-p^2)Z_p(x)^2 + \frac{1}{2}x^2Z_p'(x)^2 + c, \tag{A.1}$$

$$\begin{aligned}\int x^3Z_p(x)^2dx = \frac{1}{6}\big[&\left(x^4+p^2x^2-2p^2(p^2-1)\right)Z_p(x)^2 - 2x^3Z_p(x)Z_p'(x)\\ &+ (x^4+2(p^2-1)x^2)Z_p'(x)^2\big] + c,\end{aligned} \tag{A.2}$$

(A.3)
$$\begin{aligned}\int x^5Z_p(x)^2dx = \frac{1}{30}\big[&(3x^6+(p^2+8)x^4+4p^2(p^2-4)x^2\\ &-8p^2(p^2-1)(p^2-4))Z_p(x)^2 - 4(3x^5+2(p^2-4)x^3)Z_p(x)Z_p'(x)\\ &+(3x^6+4(p^2-4)x^4+8(p^2-1)(p^2-4)x^2)Z_p'(x)^2\big] + c.\end{aligned}$$

Of these, only (A.2) is needed for simplifying (5.2). Equation (A.1) is well known and was in fact already employed by Chiti in arriving at (5.2). We include (A.3) simply to indicate that further explicit antidifferentiation formulas can be found. In fact, one can obtain explicit formulas for $\int x^{2m+1}Z_p(x)^2dx$ for $m = 0, 1, 2, \ldots$ as well as related integrals involving $Z_p(x)Z_p'(x)$ or $Z_p'(x)^2$ via a recursive procedure.

Acknowledgments. Mark S. Ashbaugh would like to thank the Physics Department and others at the Universidad de Chile for their hospitality during his visit in April 1990 when this research was initiated. We thank Fritz Gesztesy for general comments and encouragement. In addition, it is a pleasure to thank Larry Payne and Gérard Philippin for supplying us with several of the references to the nodal line conjecture and for interesting conversations and correspondence regarding the problems considered here. Finally, we are grateful to Martin E. Muldoon for much useful information concerning Bessel function zeros, and especially to Lee Lorch for proving certain inequalities which we used above and communicating them to us prior to publication.

Note added in proof (July 19, 1993). Since completing this paper we have improved upon Marcellini's bounds for λ_3/λ_1 and $(\lambda_2+\lambda_3)/\lambda_1$ in the two-dimensional case (as alluded to in §8 above). Marcellini's bounds were 3.9170 and 5.5957, respectively, while ours are 3.9051 and 5.5249. The derivations of these bounds and other inequalities of interest will appear in our forthcoming paper, "The Range of Values of λ_2/λ_1 and λ_3/λ_1 for the Fixed Membrane Problem."

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, National Bureau of Standards Applied Mathematics Series, vol. 55, U.S. Government Printing Office, Washington, D.C., 1964.

[2] N. Anghel, *Extrinsic upper bounds for eigenvalues of Dirac-type operators*, preprint.

[3] V. I. Arnol'd, M. I. Vishik, Yu. S. Il'yashenko, A. S. Kalashnikov, V. A. Kondrat'ev, S. N. Kruzhkov, E. M. Landis, V. M. Millionshchikov, O. A. Oleinik, A. F. Filippov, and M. A. Shubin, *Some unsolved problems in the theory of differential equations and mathematical physics*, Russian Math. Surveys, 44 (1989), pp. 157–171. (Translation of Uspekhi Mat. Nauk, 44 (1989), pp. 191–202.)

[4] M. S. Ashbaugh and R. D. Benguria, *Optimal bounds for ratios of eigenvalues of one-dimensional Schrödinger operators with Dirichlet boundary conditions and positive potentials*, Commun. Math. Phys., 124 (1989), pp. 403–415.

[5] ———, *Proof of the Payne–Pólya–Weinberger conjecture*, Bull. Amer. Math. Soc., 25 (1991), pp. 19–29.

[6] ———, *A sharp bound for the ratio of the first two eigenvalues of Dirichlet Laplacians and extensions*, Ann. Mathematics, 135 (1992), pp. 601–628.

[7] ———, *Isoperimetric bound for λ_3/λ_2 for the membrane problem*, Duke Math. J., 63 (1991), pp. 333–341.

[8] ———, *Isoperimetric bounds for higher eigenvalue ratios for the n-dimensional fixed membrane problem*, Proc. Royal Soc. Edinburgh, to appear.

[9] ———, *Universal bounds for the low eigenvalues of Neumann Laplacians in n dimensions*, SIAM J. Math. Anal., 24 (1993), pp. 557–570.

[10] M. S. Ashbaugh and P. Exner, *Lower bounds to bound state energies in bent tubes*, Phys. Lett., 150A (1990), pp. 183–186.

[11] C. Bandle, *Isoperimetric Inequalities and Applications*, Pitman Monographs and Studies in Mathematics, vol. 7, Pitman, Boston, MA, 1980.

[12] J. J. A. M. Brands, *Bounds for the ratios of the first three membrane eigenvalues*, Arch. Rat. Mech. Anal., 16 (1964), pp. 265–268.

[13] Ll. G. Chambers, *An upper bound for the first zero of Bessel functions*, Math. Comp., 38 (1982), pp. 589–591.

[14] I. Chavel, *Eigenvalues in Riemannian Geometry*, Academic Press, New York, 1984.

[15] S.-Y. Cheng, *Eigenvalue comparison theorems and its geometric applications*, Math. Z., 143 (1975), pp. 289–297.

[16] G. Chiti, *Inequalities for the first three membrane eigenvalues*, Boll. Un. Mat. Ital, 18-A (1981), pp. 144–148.

[17] ———, *A bound for the ratio of the first two eigenvalues of a membrane*, SIAM J. Math. Anal., 14 (1983), pp. 1163–1167.

[18] H. L. de Vries, *On the upper bound for the ratio of the first two membrane eigenvalues*, Z. Naturforschung, 22A (1967), pp. 152–153.

[19] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, 2nd ed., Cambridge University Press, Cambridge, 1952.

[20] E. Harrell, *Some geometric bounds on eigenvalue gaps*, Comm. Partial Differential Equations, 18 (1993), pp. 179–198.

[21] J. Hersch, *On symmetric membranes and conformal radius: Some complements to Pólya's and Szegö's inequalities*, Arch. Rational Mech. Anal., 20 (1965), pp. 378–390.

[22] J. Hersch and G.-C.Rota, eds., *George Pólya: Collected Papers, Vol.* III: *Analysis*, MIT Press, Cambridge, MA, 1984.

[23] G. N. Hile and M. H. Protter, *Inequalities for eigenvalues of the Laplacian*, Indiana Univ. Math. J., 29 (1980), pp. 523–538.

[24] D. Jerison, *The first nodal line of a convex planar domain*, Int. Math. Res. Notices, 1 (1991), pp. 1–5.

[25] C.-S. Lin, *On the second eigenfunction of the Laplacian in $\mathbb{R}^2$*, Commun. Math. Phys., 111 (1987), pp. 161–166.

[26] L. Lorch, *Some inequalities for the first positive zeros of Bessel functions*, SIAM J. Math. Anal., 24 (1993), pp. 814–823.

[27] P. Marcellini, *Bounds for the third membrane eigenvalue*, J. Differential Equations, 37 (1980), pp. 438–443.

[28] W. S. Massey, *Algebraic Topology: An Introduction*, Harcourt, Brace and World, New York, 1967.

[29] A. D. Melas, *On the nodal line of the second eigenfunction of the Laplacian in $\mathbb{R}^2$*, J. Differential Geom., 35 (1992), pp. 255–263.

[30] J. R. Munkres, *Topology, A First Course*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

[31] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[32] L. E. Payne, *Isoperimetric inequalities and their applications*, SIAM Rev., 9 (1967), pp. 453–488.

[33] ———, *On two conjectures in the fixed membrane eigenvalue problem*, J. Appl. Math. Phys. (ZAMP), 24 (1973), pp. 721–729.

[34] L. E. Payne, G. Pólya, and H. F. Weinberger, *Sur le quotient de deux fréquences propres consécutives*, Comptes Rendus Acad. Sci. Paris, 241 (1955), pp. 917–919 (reprinted as pp. 410–412 of [22]).

[35] ———, *On the ratio of consecutive eigenvalues*, J. Math. Phys., 35 (1956), pp. 289–298 (reprinted as pp. 420–429 of [22] with comments by J. Hersch on pp. 521–522).

[36] G. Pólya, *On the characteristic frequencies of a symmetric membrane*, Math Z., 63 (1955), pp. 331–337 (reprinted as pp. 413–419 of [22] with comments by J. Hersch on pp. 519–521).

[37] G. Pólya and M. Schiffer, *Convexity of functionals by transplantation*, J. Analyse Math., 3 (1954), pp. 245–345 (reprinted as pp. 290–390 of [22] with comments by J. Hersch on pp. 512–515).

[38] G. Pólya and G. Szegö, *Isoperimetric Inequalites in Mathematical Physics*, Annals of Mathematics Studies, No. 27, Princeton University Press, Princeton, 1951.

[39] M. H. Protter, *Can one hear the shape of a drum? revisited*, SIAM Rev., 29 (1987), pp. 185–197.

[40] ———, *Lower bounds for the spectrum of second order operators and systems of operators*, in Maximum Principles and Eigenvalue Problems in Partial Differential Equations, P. W. Schaefer, ed., Pitman Research Notes in Mathematics Series, vol. 175, Longman Scientific and Technical, Harlow, Essex, United Kingdom, 1988, pp. 82–93.

[41] ———, *Universal inequalities for eigenvalues*, in Maximum Principles and Eigenvalue Problems in Partial Differential Equations, P. W. Schaefer, ed., Pitman Research Notes in Mathematics Series, vol. 175, Longman Scientific and Technical, Harlow, Essex, United Kingdom, 1988, pp. 111–120.

[42] C.-L. Shen, *Remarks on the second eigenvalue of a symmetric simply connected plane region*, SIAM J. Math. Anal., 19 (1988), pp. 167–171.

[43] B. Simon, *Some quantum operators with discrete spectrum but classically continuous spectrum*, Ann. Phys. (New York), 146 (1983), pp. 209–220.

[44] ———, *Nonclassical eigenvalue asymptotics*, J. Funct. Anal., 53 (1983), pp. 84–98.

[45] I. M. Singer, B. Wong, S.-T. Yau, and S. S.-T. Yau, *An estimate of the gap of the first two eigenvalues in the Schrödinger operator*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 12 (1985), pp. 319–333.

[46] E. H. Spanier, *Algebraic Topology*, McGraw-Hill, New York, 1966.

[47] C. J. Thompson, *On the ratio of consecutive eigenvalues in n-dimensions*, Stud. Appl. Math., 48 (1969), pp. 281–283.

[48] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, Cambridge, 1944.

[49] H. F. Weinberger, *An isoperimetric inequality for the n-dimensional free membrane problem*, J. Rational Mech. Anal., 5 (1956), pp. 633–636.

[50] S.-T. Yau, *Problem Section*, in Seminar on Differential Geometry, S.-T. Yau, ed., Annals of Mathematics Studies, No. 102, Princeton University Press, Princeton, 1982, pp. 669–706.

[51] ———, *Nonlinear Analysis in Geometry*, L'Enseignement Mathématique, Université de Genève, Genève, 1986.

SIAM J. MATH. ANAL.
Vol. 24, No. 6, pp. 1652–1661, November 1993

SIMULTANEOUS POLYNOMIAL APPROXIMATION*

Z. DITZIAN†, D. JIANG†, AND D. LEVIATAN‡

Abstract. A recent result (see [*Canad. J. Math.*, 44 (1992), pp. 924–940]) has bridged the gap between the Timan-type pointwise estimate and the norm estimate for polynomial approximation in $C[-1,1]$. Here, the authors show that for $f \in C^{(s)}[-1,1]$ the function and its derivatives of order j, $0 \le j \le s$, can be simultaneously approximated by polynomials of degree n and their derivatives in the appropriate manner.

Key words. simultaneous approximation, algebraic polynomial approximation

AMS subject classifications. 41A28, 41A10

1. Introduction. It is known (see [6]) that for $f \in C[-1,1]$, $0 \le \lambda \le 1$ and an integer r there exists a polynomial $P_n \in \Pi_n$, where Π_n is the class of polynomials of degree n, such that

$$|f(x) - P_n(x)| \le C\omega^r_{\varphi^\lambda}(f, n^{-1}\delta_n(x)^{1-\lambda}), \tag{1.1}$$

where $\varphi(x) = \sqrt{1-x^2}$, $\delta_n(x) = n^{-1} + \varphi(x)$, and $\omega^r_\psi(f,t)$ is given by

$$\omega^r_\psi(f,t) = \sup_{x\in I} |\Delta^r_{h\psi(x)} f(x)|, \qquad I = [-1,1]. \tag{1.2}$$

This measure of smoothness related to the step-weight function ψ was introduced and investigated in [8]. (For the case treated here $\omega^r_\psi(f,t)$ was already introduced in [5, §3].) We note that when $\lambda = 0$, (1.1) yields the Timan-type estimate for polynomial approximation and when $\lambda = 1$, (1.1) yields the norm estimate (only for $L_\infty[-1,1]$) for best polynomial approximation (see [8, Chap. 7] for the result in $L_p[-1,1]$, $1 \le p \le \infty$).

For $r = 2$ it was shown (see [7]) that $P_n \in \Pi_n$ can be found such that

$$|f(x) - P_n(x)| \le C\omega^2_{\varphi^\lambda}(f, n^{-1}\varphi(x)^{1-\lambda}). \tag{1.3}$$

In this paper we investigate the situation in which $f^{(s)} \in C[-1,1]$. We will show the following.

THEOREM 1.1. *Suppose s, r, and k are integers, $s \ge 0$, $r \ge 1$, $0 \le \lambda \le 1$ $f^{(s)}(x) \in C[-1,1]$, and $\varphi(x) = \sqrt{1-x^2}$. Then, for $n \ge s+1$ there exists $P_n \in \Pi_n$ for which*

$$|f^{(j)}(x) - P^{(j)}_n(x)| \le C\big(n^{-1}\varphi(x)\big)^{s-j} \omega^r_{\varphi^\lambda}\big(f^{(s)}, n^{-1}\delta_n(x)^{1-\lambda}\big), \qquad 0 \le j \le s, \tag{1.4}$$

and

$$|P^{(s+k)}_n(x)| \le Cn^k\delta_n(x)^{-k}\omega^r_{\varphi^\lambda}\big(f^{(s)}, n^{-1}\delta_n(x)^{1-\lambda}\big), \qquad k \ge r. \tag{1.5}$$

The construction used for Theorem 1.1 can be refined to prove the following sharper result for $r = 1, 2$.

*Received by the editors May 11, 1992; accepted for publication (in revised form) February 10, 1993.

†Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1. The work of the first author was supported by Natural Sciences and Engineering Research Council of Canada grant A4816.

‡Department of Mathematics, Tel-Aviv University, Tel-Aviv, Israel 69978. This author's work was supported by United States–Israel Binational Science Foundation (BSF) grant 8900505.

THEOREM 1.2. *Under the assumptions of Theorem* 1.1, *for $n \geq s+1$ there exists a polynomial $P_n \in \Pi_n$ such that*

$$|f^{(j)}(x) - P_n^{(j)}(x)| \leq C\big(n^{-1}\varphi(x)\big)^{s-j}\omega_{\varphi^\lambda}^{\ell}\big(f^{(s)}, n^{-1}\varphi(x)^{1-\lambda}\big) \tag{1.6}$$

for $\ell = 1, 2$ and $0 \leq j \leq s$, and

$$|P_n^{(s+k)}(x)| \leq Cn^k\delta_n(x)^{-k}\omega_{\varphi^\lambda}^{\ell}\big(f^{(s)}, n^{-1}\varphi(x)^{1-\lambda}\big) \tag{1.7}$$

for $k \geq \ell$.

Remark 1.3. In the preceding theorem it is sufficient to prove the result for $\ell = 2$ in (1.6) and (1.7) since

$$\omega_{\varphi^\lambda}^2(g,t) \leq M\omega_{\varphi^\lambda}(g,t) \quad \text{for } 0 \leq t \leq t_0 \tag{1.8}$$

(see [7, Thm. 4.1.3]). The remaining case, i.e., $k = \ell = 1$ can be treated similarly (see also the remark after (3.13)).

For the special case $\lambda = 0$ a great part (but not all) of the above theorems were proved in a series of articles spread out from the early fifties to the last decade. For $\lambda > 0$ the results are new. The earlier investigation of this type of estimate seems to go back to Timan [17] in 1951, and his work was extended in the sixties by Trigub [18] and Brudnyĭ [1] who achieved a special case of (1.4) for $\lambda = 0$. Improving the estimates by replacing $\delta_n(x)$ by $\varphi(x)$ is due to Teljakovskiĭ [16] and Gopengauz [11] who got (1.6) with $\ell = 1$ and $\lambda = 0$. Gopengauz [12] made an attempt at (1.6) for $\lambda = 0$ and other ℓ but he claimed that the result is valid for all ℓ which is known to be incorrect for $\ell \geq 3$. It seems that the first proof of (1.6) for $\ell = 2$ (under the conditions $\lambda = 0$ and $s = 0$) is due to DeVore [3], [4]. This result was further extended by Gonska and Hinnemann [10] who proved (1.6) for $\lambda = 0$, $\ell = 1, 2, \ldots$, and $0 \leq j \leq s - \ell$; and finally, independently, by Dahlhaus [2] and Li [14], who showed that for $\lambda = 0$, (1.6) holds for some $\ell = 1, 2, \ldots$ if and only if $0 \leq j \leq \min\{s-\ell+2, s\}$. In our proof of (1.6) we rely heavily on our previous construction in [7], which is valid only for $\ell = 1, 2$. Therefore, we have to limit ourselves to $\ell = 1, 2$, but we do get the full range of j, namely, $0 \leq j \leq s$. Note that we also obtain (1.7), which provides estimates on higher derivatives of the approximating polynomials. We would like to thank the referee for bringing to our attention references [2] and [14].

2. Some preliminary lemmas.

We will prove here some lemmas needed for the proof of the main result.

LEMMA 2.1. *Suppose $P_n \in \Pi_n$, $m \geq 0$, $0 \leq \lambda \leq 1$, and $\delta_n(x) = \frac{1}{n} + \sqrt{1-x^2}$ and that the increasing function $\omega(t)$ satisfies*

$$0 < \omega(\mu t) \leq M_r(\mu^r + 1)\omega(t) \tag{2.1}$$

for some positive integer r. Then

$$|P_n(x)| \leq A\big(n^{-1}\delta_n(x)\big)^m\omega\big(n^{-1}\delta_n(x)^{1-\lambda}\big), \qquad -1 \leq x \leq 1, \tag{2.2}$$

implies for $k \geq 0$

$$|P_n^{(k)}(x)| \leq C\big(n^{-1}\delta_n(x)\big)^{m-k}\omega\big(n^{-1}\delta_n(x)^{1-\lambda}\big), \qquad -1 \leq x \leq 1. \tag{2.3}$$

Proof. For $k \geq m + r(1-\lambda)$, (2.3) follows from Theorem 4.1 of [6]. For $0 < k < m + r(1-\lambda)$ we apply Lemma 2.1 of [5] and ideas from [6] or ideas from the proof of Theorem 5.3 of [15] applied to $t^m\omega(n^{-\lambda}t^{1-\lambda})$. We leave details to the reader. □

We now prove what will constitute the crucial step in the inductive proof of our main result.

LEMMA 2.2. *Suppose ℓ is an integer, $\ell \geq -1$, $0 \leq \lambda \leq 1$, $\omega(t)$ is an increasing function satisfying* (2.1) *for some integer r, $\delta_n(x) = n^{-1} + \sqrt{1-x^2}$, and g_n satisfies*

$$|g_n'(x)| \leq C\big(n^{-1}\delta_n(x)\big)^{\ell}\omega\big(n^{-1}\delta_n(x)^{1-\lambda}\big). \tag{2.4}$$

Then there exists a polynomial $Q_n \in \Pi_n$ satisfying

$$|g_n(x) - Q_n(x)| \leq M\big(n^{-1}\delta_n(x)\big)^{\ell+1}\omega\big(n^{-1}\delta_n(x)^{1-\lambda}\big) \tag{2.5}$$

and

$$|Q_n'(x)| \leq M\big(n^{-1}\delta_n(x)\big)^{\ell}\omega\big(n^{-1}\delta_n(x)^{1-\lambda}\big) \tag{2.6}$$

with $M \equiv M(C, \ell, r, \lambda)$ independent of n and g_n.

Proof. We construct the algebraic polynomial $Q_n \in \Pi_n$ by

$$Q_n(x) \equiv \int_{-\pi}^{\pi} g_n\big(\cos((\arccos\ x) - t)\big)T_n(t)dt, \tag{2.7}$$

where $T_n(t)$ is the trigonometric polynomial of order n given by

$$\begin{gathered} T_n(t) \equiv L_{[n/m],m}(t), \qquad L_{k,m}(t) = \lambda_k\left(\frac{\sin\ kt/2}{\sin\ t/2}\right)^{2m}, \\ \int_{-\pi}^{\pi} L_{k,m}(t)dt = 1 \quad \text{and} \quad m = \ell + r + 2. \end{gathered} \tag{2.8}$$

To prove (2.5) we write

$$I_n(x) = |g_n(x) - Q_n(x)| = \left| \int_{-\pi}^{\pi} \left[\int_{\cos\big((\arccos\ x) - t\big)}^{x} g_n'(u)du \right] T_n(t)dt \right|.$$

We denote the interval between $\cos\big((\arccos\ x) - t\big)$ and x by $J(x,t)$ and its length by $m\big(J(x,t)\big)$. Using straightforward computation, we have (see [8, p. 80])

$$m\big(J(x, \pm t)\big) = |x - \cos\big((\arccos\ x) \pm t\big)| \leq \Delta_{1/t}(x) \equiv |t| \ \Big(|t| + \sqrt{1-x^2}\Big). \tag{2.9}$$

Hence,

$$\begin{aligned} I_n(x) &\leq \int_{-\pi}^{\pi} \left\{ \frac{1}{m\big(J(x,t)\big)} \int_{J(x,t)} \Delta_{1/t}(x)|g_n'(u)|du \right\} T_n(t)dt \\ &\leq Cn^{-\ell} \int_{-\pi}^{\pi} \left\{ \frac{1}{m\big(J(x,t)\big)} \int_{J(x,t)} \Delta_{1/t}(x)\delta_n(u)^{\ell}\omega\big(n^{-1}\delta_n(u)^{1-\lambda}\big)du \right\} T_n(t)dt. \end{aligned}$$

Obviously,

$$\Delta_{1/t}(x) \leq |t|(n|t| + 1)\delta_n(x).$$

Using (2.1), we have

$$\omega\big(f,n^{-1}\delta_n(u)^{1-\lambda}\big)\le C_1\left(\left(\frac{\delta_n(u)}{\delta_n(x)}\right)^{r(1-\lambda)}+1\right)\omega\big(f,n^{-1}\delta_n(x)^{1-\lambda}\big).$$

Some computations following [8, Chap. 7] and the proof of Theorem 2.1 of [6] imply for $m\ge r+\ell+2$

$$I_n(x)\le C_2\big(n^{-1}\delta_n(x)\big)^{\ell+1}\omega^r\big(f,n^{-1}\delta_n(x)^{1-\lambda}\big).$$

We now follow [9, pp. 127–128] to prove (2.6). Using (2.7), $T_n(u)=T_n(-u)$, and [9, p. 128], we have

$$Q_n'(x)=-\int_0^\pi g_n'(\cos\xi)2\sin u\,T_n'(u)du,$$

with $\cos\xi\in J(x,u)$. Following [8, p. 128], we have

$$|Q_n'(x)|\le C_3\int_0^\pi |g_n'(\cos\xi)|L_{[n/m],m-1}(u)du. \tag{2.10}$$

Earlier considerations in this proof yield

$$\begin{aligned}|g_n'(\cos\xi)|&\le C\big(n^{-1}\delta_n(\cos\xi)\big)^{\ell}\omega\big(n^{-1}\delta_n(\cos\xi)^{1-\lambda}\big)\\&\le C_4\big(n^{-1}\delta_n(x)\big)^{\ell}\omega\big(n^{-1}\delta_n(x)^{1-\lambda}\big)\big((nu)^2+1\big)^{\ell}\big(((nu)^2+1)^{r(1-\lambda)}+1\big).\end{aligned}$$

Since

$$\int_{-\pi}^{\pi}|t|^\gamma L_{[n/m],m-1}(t)dt\le L_1n^{-\gamma}\quad\text{for }\gamma<2m-3,$$

the result is valid if $2\ell+2r<2m-3$, which is evident from (2.8). □

Remark 2.3. Examining the proof of Lemma 2.2, we can replace the integer ℓ satisfying $\ell\ge-1$ by any real number satisfying $\beta\ge-1$. In this case the choice $m=[\beta]+r+3$ in (2.8) will do.

LEMMA 2.4 (Trigub). *For any two integers $\ell>0$ and $m>0$ we have a polynomial $P_n\in\Pi_n$ such that*

$$\|(1-x^2)^\ell-(1-x^2)^{\ell+m}P_n(x)\|_{C[-1,1]}\le C(\ell,m)n^{-2\ell}. \tag{2.11}$$

This lemma is a corollary of Lemma 2″ of [18]. The proof in [18] is computational and long (via Lemma 2 and Lemma 2′ there). We will give a different, more accessible and, we hope, more transparent, proof below.

Proof of Lemma 2.4. For $f(x)=\sqrt{1-x^2}$ we have $\omega_\varphi(f,t)\sim t$ (see [8, p. 109(1)]). Hence (see [8, Thm. 7.2.1]), there is $Q_n\in\Pi_n$ for which

$$\left\|\sqrt{1-x^2}-Q_n(x)\right\|_{C[-1,1]}\le Cn^{-1}.$$

Writing

$$Q_{n,1}(x)=Q_n(x)-\frac{(1-x)}{2}Q_n(-1)-\frac{(1+x)}{2}Q_n(1),$$

we have $Q_{n,1}(\pm 1) = 0$, $Q_{n,1}(x) = (1-x^2)Q_{n,2}(x)$ and therefore,

$$\left\| \sqrt{1-x^2} - (1-x^2)Q_{n,2}(x) \right\|_{C[-1,1]} \le 3Cn^{-1}. \tag{2.12}$$

For $-1+n^{-2} \le x \le 1-n^{-2}$,

$$|(1-x^2)^{1/2}Q_{n,2}(x)| \le 1 + 3Cn^{-1}(1-x^2)^{-1/2} \le 1+3C.$$

Using [8, Thm. 8.4.8], we have

$$\|(1-x^2)^{1/2}Q_{n,2}(x)\|_{C[-1,1]} \le A\|(1-x^2)^{1/2}Q_{n,2}(x)\|_{C[-1+(1/n^2),1-(1/n^2)]} \le A(1+3C).$$

We now use (2.12) $2m$ times to obtain

$$\begin{aligned} &\left\| \sqrt{1-x^2} - \sqrt{1-x^2}\,(1-x^2)^m Q_{n,2}^{2m}(x) \right\| \\ &\le \left\| \sqrt{1-x^2} - (1-x^2)^m Q_{n,2}^{2m-1}(x) \right\| \\ &\quad + \left\| \left(\sqrt{1-x^2} - (1-x^2)Q_{n,2}(x)\right)(1-x^2)^{m-1/2}Q_{n,2}^{2m-1}(x) \right\| \\ &\le 3Cn^{-1}\left(1+\cdots+(A(1+3C)^{2m-1})\right). \end{aligned}$$

The desired estimate is that of $\left(\sqrt{1-x^2} - \sqrt{1-x^2}\,(1-x^2)^m Q_{n,2}^{2m}(x)\right)^{2\ell}$. The fact that the polynomial

$$P(x) = \sum_{j=1}^{2\ell} (-1)^{j-1} \binom{2\ell}{j} (1-x^2)^{m(j-1)} Q_{n,2}^{2mj}(x)$$

is not of degree n but rather of degree $4\ell m(n-1) - 2m$ does not matter since ℓ and m are fixed. □

3. The main result.

In this section we will prove Theorems 1.1 and 1.2, which together constitute our main result. It turns out that it is advantageous to prove both theorems simultaneously.

Proof of Theorems 1.1 *and* 1.2. We use Theorem 2.1 of [6] applied to $f^{(s)}(x)$ to ascertain the existence of a polynomial $q_n \in \Pi_n$ satisfying

$$|f^{(s)}(x) - q_n(x)| \le C\omega^r_{\varphi^\lambda}\left(f^{(s)}, n^{-1}\delta_n(x)^{1-\lambda}\right). \tag{3.1}$$

We also use Theorem 1.1 of [7] to obtain a polynomial $q_n(x) \in \Pi_n$ satisfying

$$|f^{(s)}(x) - q_n(x)| \le C\omega^2_{\varphi^\lambda}\left(f^{(s)}, n^{-1}\varphi(x)^{1-\lambda}\right). \tag{3.2}$$

In the construction of P_n it will make no difference whether we choose $q_n(x)$ to satisfy (3.1) or (3.2) as

$$\omega^2_{\varphi^\lambda}\left(f^{(s)}, n^{-1}\varphi(x)^{1-\lambda}\right) \le \omega^2_{\varphi^\lambda}\left(f^{(s)}, n^{-1}\delta_n(x)^{1-\lambda}\right).$$

The difference will emerge in the proofs of (1.4), (1.5), (1.6), and (1.7) as a result of that construction. Our construction will yield a polynomial P_n such that $P_n \in \Pi_{Ln}$

with some fixed integer L. This will not change the result as, in the right-hand side of $(1.4),\dots,(1.7)$, replacing n by Ln would just change the constant.

With no loss of generality we may assume that

$$f(x)=\int_0^x\int_0^{u_{s-1}}\cdots\int_0^{u_1} f^{(s)}(u)du,$$

as $f(x)$ may differ from the above only by a polynomial of degree $s-1$, which can be absorbed by $P_n(x)$.

We recall from [8, Thm. 4.1.2] that $\omega^r_{\varphi^\lambda}(f,t)$ satisfies condition (2.1). This follows from the equivalence

$$\omega^r_{\varphi^\lambda}(f,t)\sim K_{\varphi^\lambda,r}(f,t^r),$$

where

$$K_{\varphi^\lambda,r}(f,t^r)=\inf_g\left(\|f-g\|+t^r\|\varphi^{\lambda r}g^{(r)}\|\right),$$

and the basic properties of any K-functional, i.e., $K(f,t)$ is increasing and $K(f,t^r)\le a^{-r}K(f,(at)^r)$ for $0<a<1$. Hence, using Lemma 2.2 with $\ell=0$ and

$$g_n(x)\equiv\int_0^x\big(f^{(s)}(u)-q_n(u)\big)du=f^{(s-1)}(x)-\int_0^x q_n(u)du,$$

we have a polynomial $q_{n,1}(u)\in\Pi_n$ such that

$$\left|f^{(s-1)}(x)-\int_0^x q_n(u)du-q_{n,1}(x)\right|\le C_1\big(n^{-1}\delta_n(x)\big)\omega^r_{\varphi^\lambda}\big(f^{(s)},n^{-1}\delta_n(x)^{1-\lambda}\big)$$

and

$$|q'_{n,1}(x)|\le C_1\omega^r_{\varphi^\lambda}\big(f^{(s)},n^{-1}\delta(x)^{1-\lambda}\big).$$

This implies

$$|f^{(s)}(x)-q_n(x)-q'_{n,1}(x)|\le(C+C_1)\omega^r_{\varphi^\lambda}\big(f^{(s)},n^{-1}\delta_n(x)^{1-\lambda}\big).$$

We now proceed (in case $s>1$), using Lemma 2.2, with $\ell=1$ and

$$g_n(x)=f^{(s-2)}(x)-\int_0^x\int_0^{u_1}q_n(u)du\,du_1-\int_0^x q_{n,1}(u)du.$$

In general we assume by induction that for some $j<s$ we have already constructed $q_{n,0}(x),q_{n,1}(x),\dots,q_{n,j}(x)\in\Pi_n$ $\quad(q_{n,0}\equiv q_n)$ such that for $0\le i\le j$,

$$\begin{aligned}&\Bigg|f^{(s-i)}(x)-\int_0^x\int_0^{u_{i-1}}\cdots\int_0^{u_1}q_n(u)du\,du_1\cdots du_{i-1}-\cdots\\ (3.3)\qquad&\qquad-\int_0^x q_{n,i-1}(u)du-q_{n,i}(x)-\cdots-q_{n,j}^{(j-i)}(x)\Bigg|\\ &\le C\big(n^{-1}\delta_n(x)\big)^i\omega^r_{\varphi^\lambda}\big(f^{(s)},n^{-1}\delta_n(x)^{1-\lambda}\big).\end{aligned}$$

We now use Lemma 2.2 with $\ell=j$ and with $g_n(x)$ given by

$$g_n(x)=f^{(s-j-1)}(x)-\int_0^x\int_0^{u_j}\cdots\int_0^{u_1}q_n(u)du\,du_1\dots du_j-\cdots-\int_0^x q_{n,j}(u)du$$

to obtain $q_{n,j+1}$ that satisfies (3.3) for $f^{(s-j-1)}$ and $i = j + 1$. Moreover, Lemma 2.2 implies

$$|q'_{n,j+1}(x)| \le M\big(n^{-1}\delta_n(x)\big)^j \omega^r_{\varphi^\lambda}\big(f^{(s)}, n^{-1}\delta_n(x)^{1-\lambda}\big).$$

Using Lemma 2.1 with $m = j$ and $P_n(x) = q'_{n,j+1}(x)$, we have

$$(3.4) \qquad |q^{(k+1)}_{n,j+1}(x)| \le M_1\big(n^{-1}\delta_n(x)\big)^{j-k} \omega^r_{\varphi^\lambda}\big(f^{(s)}, n^{-1}\delta_n(x)^{1-\lambda}\big).$$

Hence, (3.3) is valid for $j + 1$ and $0 \le i \le j + 1$, and by induction we construct $q_n, q_{n,1}, \ldots, q_{n,s}$. We define

$$(3.5) \qquad Q_{n,0}(x) = \int_0^x \int_0^{u_{s-1}} \cdots \int_0^{u_1} q_n(x)du$$

and

$$(3.6) \qquad Q_{n,1}(x) = \int_0^x \int_0^{u_{s-2}} \cdots \int_0^{u_1} q_{n,1}(x)du\, du_1 \cdots du_{s-2} + \cdots + q_{n,s}(x).$$

The construction above yields

$$(3.7) \qquad |f^{(i)}(x) - Q^{(i)}_{n,0}(x) - Q^{(i)}_{n,1}(x)| \le A\big(n^{-1}\delta_n(x)\big)^{s-i} \omega^r_{\varphi^\lambda}\big(f^{(s)}, n^{-1}\delta_n(x)^{1-\lambda}\big)$$

for $0 \le i \le s$, $Q^{(s)}_{n,0}(x) = q_n(x)$ for $q_n(x)$ given in (3.1) or (3.2), and

$$(3.8) \qquad |Q^{(s+k)}_{n,1}(x)| \le A(k)\big(n^{-1}\delta_n(x)\big)^{-k} \omega^r_{\varphi^\lambda}\big(f^{(s)}, n^{-1}\delta_n(x)^{1-\lambda}\big) \quad \text{for } k = 0, 1, \ldots.$$

We follow Gopengauz [11] (see also [13]) with some changes and construct a polynomial $Q_{n,2}(x)$,

$$(3.9) \qquad Q_{n,2}(x) = \sum_{i=0}^{s} (1 - x^2)^i \big(\alpha_i(1 - x) + \beta_i(1 + x)\big),$$

such that

$$(3.10) \qquad Q^{(i)}_{n,2}(\pm 1) = f^{(i)}(\pm 1) - Q^{(i)}_{n,0}(\pm 1) - Q^{(i)}_{n,1}(\pm 1) \quad \text{for } 0 \le i \le s.$$

The fact that n^{-2s+2i} is increasing in i,

$$|f^{(i)}(\pm 1) - Q^{(i)}_{n,0}(\pm 1) - Q^{(i)}_{n,1}(\pm 1)| \le C n^{-2s+2i} \omega^r_{\varphi^\lambda}\big(f^{(s)}, n^{-2+\lambda}\big)$$

for $0 \le i \le s$, and induction imply

$$(3.11) \qquad |\alpha_i| + |\beta_i| \le C_1 n^{-2s+2i} \omega^r_{\varphi^\lambda}\big(f^{(s)}, n^{-2+\lambda}\big) \quad \text{for } 0 \le i \le s.$$

We now use (2.11) (Trigub's lemma) to write

$$(3.12) \qquad \|(1 - x^2)^\ell - (1 - x^2)^{\ell+m} P_{n,\ell}(x)\| \le \frac{C(m, \ell)}{n^{2\ell}}, \qquad \ell, m = 1, 2, \ldots.$$

We construct

$$(3.13)\qquad Q_{n,3}(x)=\sum_{i=1}^{s}(1-x^2)^{i+m}\big(\alpha_i(1-x)+\beta_i(1+x)\big)P_{n,i}(x)$$

with some m, $m\geq s+1$. Showing that $P_n(x)$, given by

$$P_n(x)=Q_{n,0}(x)+Q_{n,1}(x)+Q_{n,2}(x)-Q_{n,3}(x),$$

satisfies both (1.4) and (1.5) when q_n (which defines $Q_{n,0}$) is given by (3.1) and both (1.6) and (1.7) when q_n is given by (3.2) will complete the proof. (Slight modifications are needed for (1.7) and $\ell=1$, i.e., $\omega^1_{\varphi^\lambda}$ should replace $\omega^2_{\varphi^\lambda}$ in (3.2).)

We now prove Theorem 1.1. Using (3.11) and (3.12), we have

$$(3.14)\qquad \|Q_{n,3}(x)-Q_{n,2}(x)\|\leq Cn^{-2s}\omega^r_{\varphi^\lambda}(f^{(s)},n^{-2+\lambda}).$$

Using the Markov inequality, we deduce from (3.14) that

$$(3.15)\qquad \|Q^{(i)}_{n,3}(x)-Q^{(i)}_{n,2}(x)\|\leq Cn^{-2s+2i}\omega^r_{\varphi^\lambda}\big(f^{(s)},n^{-2+\lambda}\big).$$

For $-1+n^{-2}\leq x\leq 1-n^{-2}$, (1.4) is valid. For $-1\leq x\leq -1+n^{-2}$ (or $1-n^{-2}\leq x\leq 1$) we use $f^{(i)}(-1)-P^{(i)}_n(-1)=0$ for $0\leq i\leq s$ (or $f^{(i)}(+1)-P^{(i)}_n(+1)=0$) and the Taylor formula to obtain

$$(3.16)\quad f^{(i)}(x)-P^{(i)}_n(x)=\frac{(1+x)^{s-i}}{(s-i)!}\big(f^{(s)}(\xi)-P^{(s)}_n(\xi)\big)\quad\text{for } -1\leq\xi<x\leq -1+n^{-2}.$$

Hence, (3.7) and (3.15) for $i=s$ imply

$$|f^{(i)}(x)-P^{(i)}_n(x)|\leq C(1+x)^{s-i}\omega^r_{\varphi^\lambda}\big(f^{(s)},n^{-2+\lambda}\big)\quad\text{for } -1\leq x\leq -1+n^{-2},$$

which in turn implies (1.4) as $(1+x)\leq\varphi(x)/n$ for $-1\leq x\leq -1+n^{-2}$. Using the Markov–Bernstein inequality, that is,

$$|P^{(i)}_n(x)|\leq Cn^i\delta_n(x)^{-i}\|P_n\|_{C[-1,1]},\qquad P_n\in\Pi_n,$$

we now deduce from (3.14) the inequality

$$|Q^{(i)}_{n,3}(x)-Q^{(i)}_{n,2}(x)|\leq Cn^{-2s+i}\delta_n(x)^{-i}\omega^r_{\varphi^\lambda}(f^{(s)},n^{-2+\lambda}).$$

This inequality together with (3.8) implies that we have to prove (1.5) only for $Q^{(s+k)}_{n,0}=q^{(k)}_n$. For $k=r$ we recall that the estimate of $q^{(r)}_n$ is given by Theorem 6.3 of [5], that is, we have

$$(3.17)\qquad |q^{(r)}_n(x)|\leq Cn^r\delta_n(x)^{-r}\omega^r_{\varphi^\lambda}(f^{(s)},n^{-1}\delta_n(x)^{1-\lambda}).$$

The estimate of $|q^{(k)}_n(x)|$ for $k>r$ follows the same method. For $k>r$ we use the Bernstein and the Markov inequalities on the intervals $[(-1-x)/2,(1+x)/2]$, observing that

$\delta_n(u)^{-m}\omega^r_{\varphi^\lambda}(f^{(s)}, n^{-1}\delta_n(u)^{1-\lambda})$ is bounded by $C(m)\delta_n(x)^{-m}\omega^r_{\varphi^\lambda}(f^{(s)}, n^{-1}\delta_n(x)^{1-\lambda})$ for u in that interval.

For the proof of (1.6) we may use (3.2) and hence $f^{(s)}(\pm 1) = q_n(\pm 1)$. Furthermore, $Q^{(s)}_{n,3}(\pm 1) = 0$, and therefore (3.10) reduces to

$$Q^{(s)}_{n,1}(\pm 1) + Q^{(s)}_{n,2}(\pm 1) = 0. \tag{3.18}$$

We can now prove the inequality (1.6). Note that for $-1 + n^{-2} < x < 1 - n^{-2}$, (1.6) has already been proved as in this case; there is no difference between (1.4) and (1.6) but the constant. For $-1 \le x \le -1 + n^{-2}$ (or $1 - n^{-2} < x \le 1$) we use both (3.16) and (3.17) and write

$$\begin{aligned} f^{(i)}(x) - P^{(i)}_n(x) &= \frac{(1+x)^{s-i}}{(s-i)!}\left(f^{(s)}(\xi) - P^{(s)}_n(\xi)\right) \\ &= \frac{(1+x)^{s-i}}{(s-i)!}\left(f^{(s)}(\xi) - q_n(\xi)\right) \\ &\quad + \frac{(1+x)^{s-i}}{(s-i)!}(1+\xi)\big(- Q^{(s+1)}_{n,1}(\eta) \\ &\qquad\qquad - Q^{(s+1)}_{n,2}(\eta) + Q^{(s+1)}_{n,3}(\eta)\big) = I_1 + I_2 \end{aligned}$$

for $-1 < \eta < \xi < x \le -1 + n^{-2}$. Using (3.2) and $\varphi(\xi) \le \varphi(x)$, we conclude that I_1 is smaller than the right-hand side of (1.6). Using (3.8) for $k = 1$ and (3.15) for $i = s + 1$, we have

$$\begin{aligned} |I_2| &\le C(1+x)^{s-i+1} n^2 \omega^2_{\varphi^\lambda}(f^{(s)}, n^{-2+\lambda}) \\ &\le C\left(n^{-1}\varphi(x)\right)^{s-i}(1+x) n^2 \omega^2_{\varphi^\lambda}\left(f^{(s)}, n^{-2+\lambda}\right). \end{aligned}$$

We now note that for any F, $\omega^r_{\varphi^\lambda}(F, t)$ satisfies

$$\omega^2_{\varphi^\lambda}(F, t) \le C_1 a^{-2} \omega^2_{\varphi^\lambda}(F, at)$$

for $a < 1$. Setting $F = f^{(s)}$, $t = n^{-2+\lambda}$, and $a = \left(n\varphi(x)\right)^{1-\lambda}$, we have

$$\begin{aligned} |I_2| &\le C_1\left(n^{-1}\varphi(x)\right)^{s-i}(1+x) n^2 \left(n\varphi(x)\right)^{-2+2\lambda} \omega^2_{\varphi^\lambda}\left(f^{(s)}, n^{-1}\varphi(x)^{1-\lambda}\right) \\ &\le C_1\left(n^{-1}\varphi(x)\right)^{s-i} \omega^2_{\varphi^\lambda}\left(f^{(s)}, n^{-1}\varphi(x)^{1-\lambda}\right). \end{aligned}$$

This completes the proof of (1.6). The estimate (1.7) follows that of Theorem 6.3 of [6] almost word for word. □

REFERENCES

[1] JU. A. BRUDNYĬ, *Generalizations of a theorem of A. F. Timan*, Dokl. Akad. Nauk USSR, 148 (1963), pp. 1237–1240; Soviet Math. Dokl., 4 (1963), pp. 244–247. (In English.)

[2] R. DAHLHAUS, *Pointwise approximation by algebraic polynomials*, J. Approx. Theory, 57 (1989), pp. 272–277.

[3] R. A. DeVore, *Pointwise approximation by polynomials and splines*, Theory of Approximation of Functions, S. B. Stečkin and S. A. Teljakovskiĭ, eds., Proc. Intern. Conf., Kaluga, 1975, Nauka, Moskow, 1977, pp. 132–141.

[4] ———, *Degree of approximation*, Approximation Theory II, G. G. Lorentz, C. K. Chui and L. L. Schumaker, eds., Proc. Internat. Sympos., Austin, TX, 1976, Academic Press, New York, 1976, pp. 117–161.

[5] Z. Ditzian, *On interpolation of $L_p[a, b]$ and weighted Sobolev spaces*, Pacific J. Math., 90 (1980), pp. 307–323.

[6] Z. Ditzian and D. Jiang, *Approximation by polynomials in $C[-1, 1]$*, Canad. J. Math., 44 (1992), pp. 924–940.

[7] Z. Ditzian, D. Jiang, and D. Leviatan, *Shape preserving polynomial approximation in $C[-1, 1]$*, Proc. Camb. Philos. Soc., 112 (1992), pp. 309–316.

[8] Z. Ditzian and V. Totik, *Moduli of Smoothness*, Springer-Verlag, New York, 1987.

[9] L. O. Ferguson, *Approximation by Polynomials with Integral Coefficients*, American Mathematical Society, Providence, RI, 1980.

[10] H. Gonska and E. Hinnemann, *Punktweise abschätzungen zur approximation durch algeraische polymone*, Acta Math. Hungar., 46 (1985), pp. 243–254.

[11] I. E. Gopengauz, *A theorem of A.F. Timan on the approximation of functions by polynomials on a finite segment*, Mat. Zametki, 1 (1967), pp. 163–172. (In Russian.) Math. Notes, 1 (1967), pp. 110–116. (In English.)

[12] ———, *A question about the approximation of functions on a finite segment and in a domain with corners*, Teor. Funkciĭ Funkcional. Anal. i Prilozhen., 4 (1967), pp. 204–210. (In Russian.)

[13] T. Kilgore, *An elementary simultaneous approximation theorem*, Proc. Amer. Math. Soc., to appear.

[14] W. Li, *On Timan-type theorems in algebraic polynomial approximation*, Acta Math. Sinica, 29 (1986), pp. 544–549.

[15] G. G. Lorentz, *Approximation of Functions*, Holt, Rinehart and Winston, New York, 1966.

[16] S. A. Teljakovskiĭ, *Two theorems on the approximation of functions by algebraic polynomials*, Mat. Sb., 70 (1966), pp. 252–265. (In Russian.)

[17] A. F. Timan, *Strengthening of Jackson's theorem on the approximation of continuous functions on a finite segment of the real axis*, Dokl. Akad. Nauk USSR, 78 (1951), pp. 17–20. (In Russian.)

[18] R. M. Trigub, *Approximation of functions by polynomials with integral coefficients*, Iz. Akad. Nauk SSSR Ser. Mat., 26 (1962), pp. 261–280. (In Russian.)

SIAM J. MATH. ANAL.
Vol. 24, No. 6, pp. 1662–1665, November 1993

013

A NOTE ON THE GENERALISED KATZENELSON'S METHOD FOR PIECEWISE LINEAR RESISTIVE NETWORKS*

V. C. PRASAD† AND P. B. L. GAUR‡

Abstract. Ohtsuki, Fijisawa, and Kumagai [*SIAM J. Math. Anal.*, 8 (1977), pp. 69–99] proposed a generalisation of Katzenelson's method that is useful for a large class of piecewise linear resistive networks. In this note, errors are pointed out and the correct proof is presented.

Key words. generalised Katzenelson's method, piecewise linear resistive networks, nonlinear equations, numerical methods

1. Introduction.

Katzenelson [1] proposed an algorithm which is guaranteed to converge to a solution starting from any initial point, provided that the corresponding piecewise linear resistive network has a unique solution. In a significant paper, Ohtsuki, Fujisawa, and Kumagai [3] extended this method to a large class of piecewise linear resistive networks, particularly those containing diodes, especially tunnel diodes, Ebers–Moll transistors, etc. Their method chooses an initial point satisfying what is called the "1-degree condition." Starting from such an initial point, they showed that their generalisation of the Katznelson's method (they called it the "generalised Katzendson's method") is guaranteed to converge to a solution. It is the purpose of this paper to correct some errors in their paper.

2. Deficiencies in the proof of the generalised Katzenelson's method.

The reader is referred to [3] for notation and other details not given here. Let $x^{(o)}$ be an initial point satisfying the 1-degree condition. (This essentially means that $x^{(o)}$ is the only solution of the $y^{(o)} \triangleq f(x^{(o)})$, where $f(x) = y$, $x \in R^n$, $y \in R^n$ and f is a continuous piecewise linear function.) Let $y^{(f)}$ be a point whose solution is required. Starting from $x^{(o)}$, trace the inverse image (i.e., the solution curve L^{-1}), of the line L, joining $y^{(o)}$ and $y^{(f)}$ using "generalised Katzenelson's method" [3]. In Theorem 6 of [3] Ohtsuki et al. stated that this method always converges to a solution. The proof of this theorem is wrong as explained in the following.

(1) Whenever the solution curve hits a corner point, it may or may not be extendable into a new region. To overcome this, they used Theorem 2 (page 81, last line of [3]). In the proof of Theorem 2 they stated that "The degree preserves a constant value, say d." This is incorrect. It is possible to show through an example that even though f is norm-coercive, the restriction of this function at a corner point may not be norm-coercive [4]. Consequently, although the degree is a constant value globally, it is not so locally at a corner point. Thus the assumption that the degree is a constant locally in the proof of their Theorem 2 is wrong. Since Theorem 2 is incorrect, it cannot be used in the proof of Theorem 6.

(2) They have not considered the following possibilities.

(a) In Fig. 1(a), the solution curve ends abruptly at a corner point. Even if their Theorem 2 is assumed to be correct, their Remark 8 cannot guarantee that the

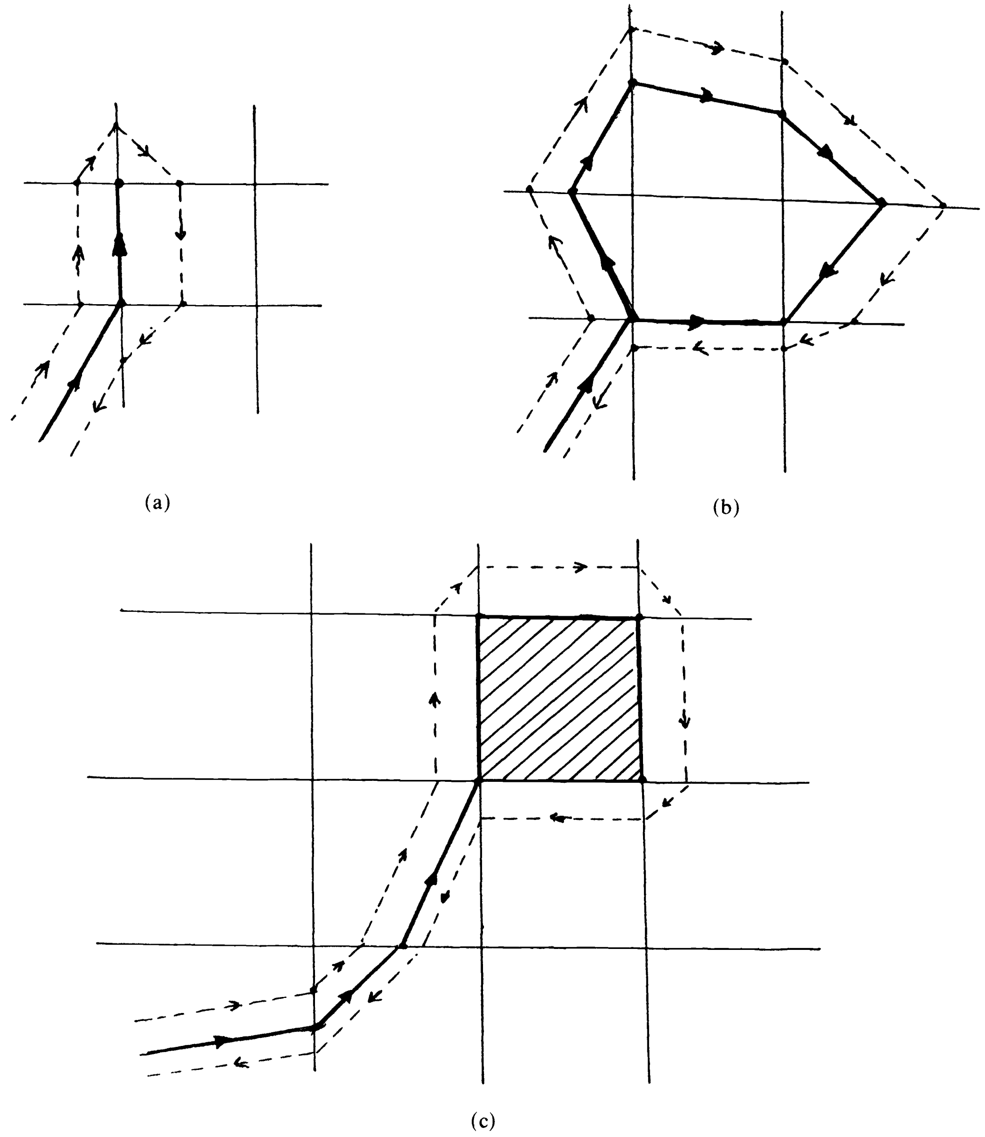
* Received by the editors April 13, 1992; accepted for publication (in revised form) February 17, 1993.

† Department of Electrical Engineering, Indian Institute of Technology, Hauz Khas, New Delhi-110016, India.

‡ Department of Electrical Engineering, Dayalbagh Educational Institute, Dayalbagh, AGRA-282005, India.

solution curve can be extended into a new region. (They used Remark 8 on page 82 to prove Theorem 6.)

(b) The solution curve shown in Fig. 1(b) satisfies their Remark 8. Similarly, Fig. 1(c) is another possibility. But their arguments do not consider these situations.


(a)

(b)

(c)

FIG. 1. *Some theoretically possible situations of the solution curve. The solution curve is shown in thick lines. Its perturbation is shown in dotted lines.*

3. Correct proof of Theorem 6. The above situations do not occur if the initial point $x^{(o)}$ satisfies the 1-degree condition and an additional condition (to be explained later) is satisfied. This can be proved using the perturbation technique of Fujisawa and Kuh [2]. For this purpose we need the following definitions. A point, $Z \in R^n$, is said to be a point at infinity if and only if there exists at least one $i, 1 \leq i \leq n$, such that Z_i is arbitrarily large. The line segment of a solution curve in a region R is said to be an end segment if there is a point at infinity on the line segment.

Let $L : y = y^{(o)} + \lambda(y^{(f)} - y^{(o)}, -\infty < \lambda < \infty$ be the doubly infinite line joining $y^{(o)}$ to $y^{(f)}$. Trace the solution curve starting from $x^{(o)}$. Let $x^{(o)}$ lie in a region R_1. If L^{-1} lies entirely in R_1, then a solution of $y^{(f)}$ is obtained anyway and there is nothing to prove (case A of [3]). The solution curve being traced from $x^{(o)}$ in R_1 may eventually come back to $x^{(o)}$, i.e., the solution curve containing $x^{(o)}$ is a loop. This is not possible if $x^{(o)}$ satisfies the 1-degree condition (case B of [3]). This leaves us with only one possibility.

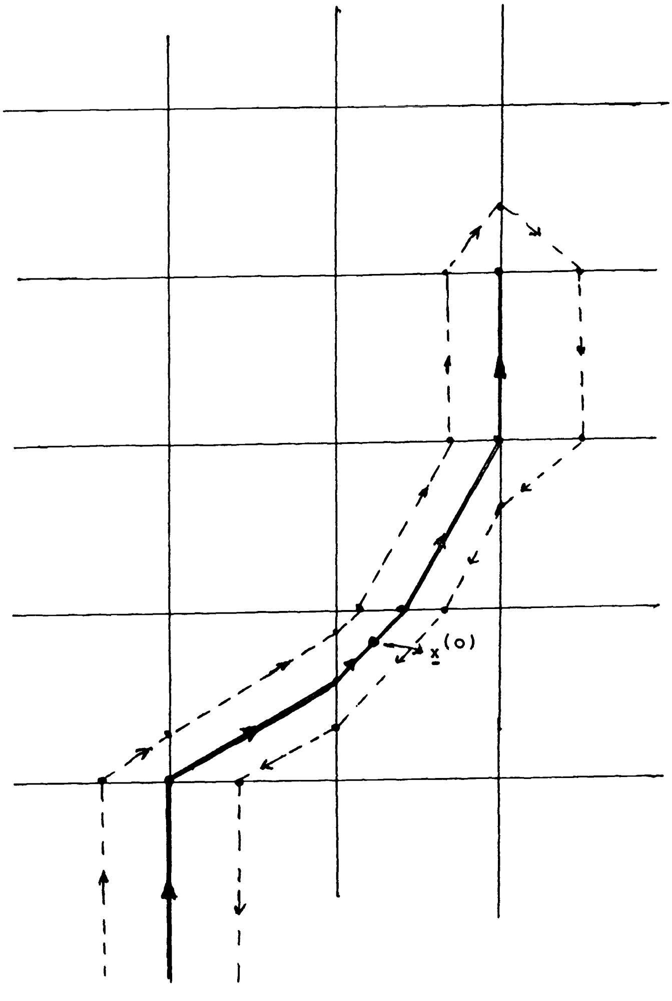


FIG. 2. *A situation in which the generalised Katzenelson's method may not give a solution: $x^{(o)}$ satisfies 1-degree condition.*

The solution curve being traced from $x^{(o)}$ has one end segment in some regions but does not eventually terminate in another unbounded region having another end segment. (The possibilities in Fig. 1 depict this.) In such a case L is perturbed. As explained in Fujisawa and Kuh [2], it is always possible to find a perturbation of L such that the perturbed solution curve does not hit a corner point. Therefore the perturbed solution curve must have two end segments. If the original solution curve has only one end segment (as $\lambda \to -\infty$), say, in the interior of the region R, then the two end segments of the perturbed solution curve also lie in R. If R contains $x^{(o)}$, then its Jacobian is nonsingular. Therefore there cannot be two parallel line segments in a region mapped on to the same line segment in the y-space. Even if $x^{(o)}$ does not lie in R, this statement is still true. This is because the Jacobian of R must be nonsingular as f is norm-coercive.

As $\lambda \to -\infty$, if the end segment of the solution curve starting from $x^{(o)}$ does not lie in the interior of a region, then the generalised Katzenelson's method may not give a solution, at least in theory. Fig. 2 depicts such a possibility. From a practical point of view, since every point in a small open neighborhood of a point satisfying the 1-degree condition also satisfies the 1-degree condition, it is, in general, possible to choose an initial point in that neighborhood such that as $\lambda \to -\infty$, the end segment of L^{-1} lies in the interior of a region. Thus Fig. 2, even if it is true, can be avoided in practice.

REFERENCES

[1] J. KATZENELSON, *An algorithm for solving resistive networks*, Bell Syst. Tech. J., 44 (1965), pp. 1605–1620.

[2] T. FUJISAWA AND E. S. KUH, *Piecewise linear theory of nonlinear networks*, SIAM J. Appl. Math., 22 (1972), pp. 307–328.

[3] T. OHTSUKI, T. FUJISAWA, AND S. KUMAGAI, *Existence theorems and a solution algorithm for piecewise linear resistor networks*, SIAM J. Math. Anal., 8 (1977), pp. 69–99.

[4] P. B. L. GAUR, *Existence, uniqueness and determination of solutions of piecewise linear resistive networks*, Ph.D. thesis, Dept. of Electrical Engineering, Indian School of Mines, Dhanbad, India, 1984.